RISE Working Paper 21/084
November 2021

Preparation, Practice, and Beliefs: A Machine Learning Approach to Understanding Teacher Effectiveness

Deon Filmer, Vatsal Nahata, and Shwetlena Sabarwal

Abstract

This paper uses machine learning methods to identify key predictors of teacher effectiveness, proxied by student learning gains linked to a teacher over an academic year. Conditional inference forests and the least absolute shrinkage and selection operator are applied to matched student-teacher data for Math and Kiswahili from Grades 2 and 3 in 392 schools across Tanzania. These two machine learning methods produce consistent results and outperform standard ordinary least squares in out-of-sample prediction by 14-24 percent. As in previous research, commonly used teacher covariates like teacher gender, education, experience, and so forth are not good predictors of teacher effectiveness. Instead, teacher practice (what teachers do, measured through classroom observations and student surveys) and teacher beliefs (measured through teacher surveys) emerge as much more important. Overall, teacher covariates are stronger predictors of teacher effectiveness in Math than in Kiswahili. Teacher beliefs that they can help disadvantaged and struggling students learn (for Math) and they have good relationships within schools (for Kiswahili), teacher practice of providing written feedback and reviewing key concepts at the end of class (for Math), and spending extra time with struggling students (for Kiswahili) are highly predictive of teacher effectiveness, as is teacher preparation on how to teach foundational topics (for both Math and Kiswahili). These results demonstrate the need to pay more systematic attention to teacher preparation, practice, and beliefs in teacher research and policy.

Keywords: education; teacher performance; teacher value-added; teacher mindsets; student achievement
JEL Codes: I20; I21; I25; I28; J45

Preparation, Practice, and Beliefs: A Machine Learning Approach to Understanding Teacher Effectiveness

Deon Filmer
The World Bank

Vatsal Nahata
The World Bank

Shwetlena Sabarwal
The World Bank

Acknowledgements:
We would like to thank the Research on Improving Systems of Education (RISE) program for funding and to the RISE Tanzania team for background work and inputs. Comments and guidance from Samer Al-Samarrai, Noam Angrist, Marina Bassi, Paolo Brunori, Jacobus Cilliers, Xiaoyan Liang, Daniel Mahler, Chiara Masci, Halsey Rogers, Dario Sansone, Fritz Schiltz, Jan Spiess, Falco Stoffi, Inaam Ul Haq, and the RISE Quality Assurance Team are gratefully acknowledged. Diwakar Kishore provided excellent research assistance.

The authors of this paper are listed alphabetically.

This is one of a series of working papers from "RISE"—the large-scale education systems research programme supported by funding from the United Kingdom's Foreign, Commonwealth and Development Office (FCDO), the Australian Government's Department of Foreign Affairs and Trade (DFAT), and the Bill and Melinda Gates Foundation. The Programme is managed and implemented through a partnership between Oxford Policy Management and the Blavatnik School of Government at the University of Oxford.

Please cite this paper as:
Filmer, D., Nahata, V. and Sabarwal, S. 2021. Preparation, Practice, and Beliefs: A Machine Learning Approach to Understanding Teacher Effectiveness. RISE Working Paper Series. 21/084.
https://doi.org/10.35489/BSG-RISE-WP_2021/084

1. Introduction

There is strong agreement that teachers matter a lot for student learning;[1] but little agreement on which specific teacher factors matter most. Studies document wide variation in teacher effectiveness that is not well explained by observable teacher characteristics (e.g., McCaffrey et. al 2004; Jacob and Lefgren, 2005; Rivkin, Hanushek, and Kain, 2005; Gordon, Kane, and Staiger, 2006; Kane, Rockoff, and Staiger, 2008). Specifically, observable and widely available teacher characteristics such as teacher qualifications, test scores, training, and experience appear to be weak predictors of teacher contributions to student learning in high-income countries (Rockoff 2004; Rivkin, Hanushek and Kain 2005; Aaronson, Barrow and Sander 2007, Staiger and Rockoff 2010). This finding is mirrored in recent studies from low- and middle-income countries. Research from Pakistan and India does not find a strong relationship between teacher qualifications and teacher value-added in either government or private schools (Bau and Das 2020; Azam and Kingdon 2015); in Ecuador teacher entry-exam performance explains a small fraction of the variation in student learning (Cruz-Aguayo et al. 2017). Bau and Das (2020) find in Pakistan's context that observed teacher characteristics account for less than 5 percent of the variation in teacher value-added.

Using machine learning methods on a rich student-teacher data set from Tanzania, this paper identifies key predictors of teacher effectiveness, proxied through student learning gains linked to a teacher over an academic year, closely linked to the concept of teacher value-added (TVA).[2] Specifically, it explores which aspects of a teacher – who teachers are, what teachers know, what teachers do, or what teachers believe – are most predictive of student learning gains.

Machine learning (ML) is a well-suited (albeit novel) approach for identifying key predictors of teacher effectiveness from a large set of teacher, student, and school covariates since most studies in the TVA literature use linear modeling techniques (Koedel et al. 2015). ML applications, which are increasingly common in econometrics (Athey and Imbens 2016, Mullainathan and Spiess 2017), often involve predictions about some variables given others. They manage to uncover generalizable patterns and discover complex structures that were not specified in advance (Mullainathan and Spiess 2017). ML algorithms attempt to select flexible models that fit the data well, but not so well that out-of-sample prediction is compromised (Athey and Imbens 2016). They can be particularly successful on high-dimensional data where we observe many pieces of information on each unit (Athey and Imbens 2016).

In our case where data are relatively high dimensional (with 52 explanatory variables), using ML algorithms helps us avoid ad-hoc model selection. Instead, the ML algorithm helps show which teacher covariates matter more for predicting student learning gains, by allowing for highly flexible models that are evaluated on the basis of out-of-sample replicability. ML algorithms also help avoid the problem of multicollinearity

[1] See for example: Hanushek and Rivkin 2010; Nye et al. 2004; Chetty et al. 2014a; 2014b; Buhl-Wiggers et al. 2017; Bau and Das, 2020.

[2] Several studies show that teacher value-added measures, which control for a student's prior-year test scores, provide unbiased forecasts of teachers' causal impacts on student achievement (Bacher-Hicks et al 2017, Glazerman and Protik 2015, Chetty et al. 2014, Bacher-Hicks, Kane, and Staiger 2014, Rothstein 2014).

(Dormann et al. 2013). We use two ML algorithms Conditional Inference Forests (CIF) and Least Absolute Shrinkage and Selection Operator (LASSO).

We find that ML methods outperform standard OLS in out-of-sample prediction by 14-24 percent. Also, the identified variables of importance are largely consistent across our two, very different, ML methods. ML results are in line with previous research in that commonly used teacher characteristics such as, teacher gender, education qualifications, experience etc. do not seem to hold much predictive power for student learning gains. Instead, what teachers do in terms of specific classroom practices (measured through classroom observations and student surveys) and what teachers believe in terms of how they perceive the abilities of their students and the environment around them (measured through teacher surveys) are consistently revealed to be important.

Overall, teacher covariates matter more, and differently, for Math than Kiswahili. For Math, the teacher belief that they can help disadvantaged and struggling students learn; the teacher practice of providing clear and helpful written feedback (on homework and tests); and the teacher preparation in teaching foundational concepts are the three most predictive factors for student learning gains. For Kiswahili, where teacher (and other observable) covariates are on the whole less predictive, teacher preparation, practice, and beliefs still emerge as being important. Specifically, the teacher belief that action is taken against poor teacher performance; the teacher practice of providing extra help to struggling students[3]; and (as in Math) teacher preparation in teaching foundational concepts are the three most predictive teacher covariates for student learning gains.

Our paper contributes in two ways to the literature on teacher effectiveness. It demonstrates how machine learning techniques can help address long-standing prediction problems in education economics, including that of predicting student learning gains linked to a teacher. Applications of ML in education economics are still not very common. We were able to find three types– one for predicting student dropouts (Aulck et al. 2016; Adelman et al. 2018; Sansone 2019), the second for predicting student performance in international tests like TIMSS and PISA and national tests (Agasisti et al. 2018) and finally in modeling teacher productivity (Chalfin et al. 2016). By applying ML techniques to a rich set of control variables, we are able to explore the question of teacher effectiveness with much more granularity, while letting the data speak. This last part is particularly important because it allows for the use of flexible and non-parametric approaches in estimation while restricting arbitrary judgements on the part of the researcher, the scope for which only increases with a richer set of controls.

Another set of important but less overtly actionable insights relate to the importance of teacher beliefs in determining teacher effectiveness. The paper shows that teacher beliefs about whether students can learn and whether they have good relationships matter for their effectiveness. This finding corroborates a sizeable but scattered body of evidence on the importance of teacher beliefs for student outcomes (Sabarwal et. al 2021). Given that teacher beliefs have not been given much systematic attention in the design and implementation of teacher policies, these findings suggest that these beliefs might be an important but missing ingredient of programs and policies for teacher effectiveness. The question is – can these beliefs be changed through interventions and policies? Recent research from different disciplines show that they can. A body of research in education (e.g., Dweck 2006, Yeager et al. 2012, Paunesku et al. 2015) and

[3] During breaks, lunch, or after school.

organizational psychology (e.g. Heslin, Latham and VandeWalle 2005) have revealed how fixed mindsets can be shifted and how this can help improve motivation and performance. This paper also highlights the importance of further exploring this line of work.

Our paper is organized as follows. Section 2 provides information on data, estimation strategy (including a conceptual introduction to the CIF and LASSO algorithms), and limitations; Section 3 provides a descriptive analysis of teacher-level and other covariates, Section 4 presents the main results, and Section 5 concludes.

2. Data, Methodology, and Limitations

2.1 Data

This study is a part of the Research on Improving Systems of Education (RISE) program for Tanzania, wherein several researchers are using the same data for different studies looking at various aspects of the Tanzanian education system and reform.[4] The data for this paper comes from 392 schools randomly sampled from 392 wards across 22 districts in 6 representative regions across Tanzania. Our final sample includes 436 teachers and 3,019 students. The baseline survey was conducted between February-May 2019 and the follow-up survey between January-April 2020, targeting 748 teachers and 6,586 students from Grades 2 and 3.

Three instruments were used to collect data on teacher covariates. These include a detailed teacher survey which also has a dedicated module on teacher mindsets; a teacher subject content knowledge assessment; and classroom observation of teachers using the *Teach Classroom Observation* tool.[6]

For the detailed teacher survey, 10 teachers were randomly selected from the complete teacher roster for the school provided by the head-teacher. To the extent possible, the survey was targeted at teachers teaching Math and Kiswahili in Grades 2 and 3. After the survey, only the Grade 2 and 3 teachers were invited to take the teacher assessment. Teacher assessments were subject-based and linked to the curriculum. For Kiswahili, teachers were expected to read a short text and answer 8 comprehension questions, while for Math they answered 10 questions about basic algebra operations and geometry.

Finally, in each school, one Grade 2 and one Grade 3 teacher were randomly selected for classroom observation using the *Teach* Classroom Observation tool (Molina et. al. 2018). *Teach* allows enumerators to rate teaching practices through two 15-minute observations during a lesson. The practices are organized into nine dimensions: Supportive Learning Environment, Positive Behavioral Expectations, Lesson Facilitation, Checks for Understanding, Feedback, Critical Thinking, Autonomy, Perseverance and Socio-Emotional Skills. These dimensions are measured on a five-point scale and then averaged across the two 15-minute observations.

The study also included a student survey and student assessment. For each school, around 20 students (10 each from Grades 2 and Grade 3) were randomly selected from a list of all Grade 2 and 3 students (provided by the head-teacher). Students were tested on foundational concepts in Math and Kiswahili. These tests

[4] For more details see: https://riseprogramme.org/countries/tanzania
[6] https://www.worldbank.org/en/topic/education/brief/teach-related-blogs

were developed by Tanzania education professionals and are similar to the Uwezo annual learning assessment – a nationwide assessment used to measure learning in Tanzania (see Mbiti et al. 2021 for exact test creation). The Math test focused on counting, basic addition, subtraction, multiplication and division, while the Kiswahili test focused on correctly reading words, writing sentences and comprehension.[5] For calculating student learning gains, the same set of test questions were used at baseline and follow-up. The test was of a slightly higher level for Grade 3 compared to Grade 2 students. The tests were low stakes and designed to test a range of abilities such that scores could be equated across years using a set of linked questions in baseline and follow up. These features allow us to test children on the same knowledge scale. A student survey was also administered to collect data on student characteristics and student perceptions about teacher practices (e.g. practices that teachers did or did not engage in with students).

Our data set contains 52 explanatory variables that can be divided into the following categories: (i) student-level variables such as age, household asset ownership, baseline score and whether they attended private tuitions for the particular subject; (ii) school-level variables such as the pupil teacher ratio at the school, whether the school is in an urban or rural location and certain institutional/governance variables; and (iii) teacher level variables. We divide our teacher level variables into 4 categories: (i) Who teachers are (teacher characteristics); (ii) What teachers know (teacher knowledge measured through the teacher assessment); (iii) What teachers do (teacher practice measured through teacher classroom observation); and (iv) What teachers believe (teacher mindsets). These are discussed further in Section 3 and Annex 1.

The unit of observation is the teacher and the outcome of interest is average student learning gains between baseline and follow-up (approximately 8 months) for the teacher. We construct estimates of student learning gains using the matched teacher-student database and student assessment data from baseline and follow-up. Student learning gains linked to a particular teacher are calculated as the percentage correct score in the follow-up student assessment minus the percentage correct score in the baseline student assessment, averaged across their students. We then model student learning gains using our ML algorithms on a host of student, school and teacher level covariates. We conduct the analysis separately for Math and Kiswahili.

The estimation of student learning gains for a teacher can be seen as analogous to the estimation of Teacher Value Added (TVA), however there are some differences between our estimation of student learning gains and the way in which TVA is often estimated in the standard education economics literature. In this literature, TVA is estimated as the teacher fixed effect from a regression of student follow-up test scores on student level covariates including lagged test scores (see Koedel et al. 2015 for a comprehensive review)—this is often referred to as step 1. This teacher fixed effect is then regressed on a host of teacher and school level covariates to find out teacher characteristics that best predict TVA (Rockoff 2004, Chetty et al. 2014a, Koedel et al. 2015)—referred to as step 2. We choose the student learning gain approach over the standard TVA approach to avoid imposing a linear functional form in either step 1 or step 2. This allows

[5] For Grade 2, the Math portion of the test had 12 questions while the Kiswahili portion had 16 questions. For Grade 3, The Math portion had 17 questions while the Kiswahili portion had 15 questions (to ensure uniformity in comparison, we only chose those questions that were repeated in baseline and follow-up).

the Machine Learning algorithms maximum room to use highly flexible and interactive functional forms in a manner that is completely driven by the underlying data.[6]

There are two sources of attrition in our data. First, some students could not be contacted at follow-up and second, several teachers could not be matched to students.[7] Ultimately, we were able to map 3,019 students to 436 teachers. Since we conduct our analysis based on the subject taught by a given teacher, our final sample for analysis includes 346 Math teachers matched with 2,359 students; and 336 Kiswahili teachers matched with 2,297 students. In the Table 1, we compare teacher characteristics at baseline in the full sample and the final sample. This comparison suggests that attrition of teachers is mostly random on observables.

2.2 Methodology

Our main goal is to identify which covariates matter most for predicting student learning gains. To do this, we rely on machine learning (ML) approaches. In this section we provide a brief overview of our overall approach and of our chosen ML algorithms.

We use two machine learning algorithms: Conditional Inference Forests (CIF) and Least Absolute Shrinkage and Selection Operator (LASSO) to predict student learning gains. CIF and LASSO are both supervised algorithms where we have data on the dependent variable. The goal of supervised learning is to learn a function that, given a sample of dependent and explanatory variables, best approximates the relationship between them. Given the availability of data on the dependent variable, the supervised algorithm can compare its estimates to the actual values of the dependent variable.[8] Typically, data are split into two sets: a training set and a test set. The algorithm learns about the relationship between the dependent and explanatory variables using the training set. The test set is not used by the algorithm during the model building process and is therefore used to empirically evaluate its out of sample performance (details in Annex 2).

Within the family of supervised ML algorithms, we chose CIF and LASSO because (i) they approach the variable selection problem in different ways, with CIF belonging to the non-parametric and LASSO to the parametric class of ML models; (ii) their suitability for variable selection in high-dimensional data like ours; and (iii) their growing popularity in economics and the broader social science literature (Varian 2014; Mullainathan & Spiess 2017).

We benchmark the predictive performance of our ML models to the standard linear regression model used in the extant literature, that is Ordinary Least Squares (OLS). Next, we show the key variables of importance for predicting student learning gains identified by the ML methods. Finally, we use the ML-

[6] For the sake of robustness, we also estimate TVA using the traditional approach by first calculating teacher fixed effects and subsequently model these teacher fixed effects using our ML algorithms on teacher level covariates. As seen in Annex 3, our main results remain robust in this more traditional TVA specification.

[7] The official mapping of students and teachers at the school level happens via "streams" while the actual mapping of students and teachers is done via informal "groups" which comprise students from multiple grades and subjects in a single class. This system of "groups" is not documented at the school level.

[8] Unsupervised learning algorithms, on the other hand, do not have data on the dependent variable, so their goal is to infer the natural structure present within a set of explanatory variables. An example is Principal Components Analysis or Clustering data based on a given set of covariates.

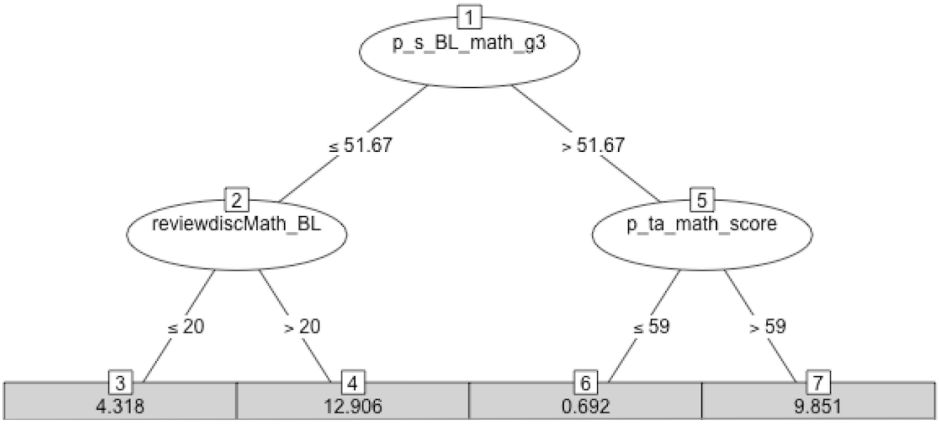
identified variables to run a parsimonious OLS regression for student learning gains. We do this to further analyze the relative importance of our ML-identified variables. In sections 2.2.1 and 2.2.2 we provide a conceptual introduction to the CIF and LASSO models. In section 2.2.3 present the OLS model which we use to benchmark the performance of our ML models.

2.2.1 Conditional Inference Trees & Forests (CIF)

Trees or Decision Trees divide the covariate space (X_1,X_2,…X_k) into M mutually exclusive regions/groups (G_1,G_2,…G_m) using a well-defined splitting criterion. This implies that every observation finds itself as part of any one group, with each group being homogenous in the expression of some variables in the covariate space. For any observation y_i that finds itself in a given group G_m, the decision tree simply predicts $\hat{y}_i$ to be the mean y value of all observations that find themselves in the same group. Due to their inherent non-parametric structure, trees are able to accommodate flexible and highly interactive relationships between the explanatory and dependent variables.

The precise manner in which splits are made depends on the variant of the tree used. In this paper we use conditional inference trees (CIT) proposed by Hothorn et al. (2006) instead of the standard regression tree (see Loh 2011 for an introduction) because the latter are biased towards selecting continuous variables with more split points as compared to categorical variables (Hothorn et al. 2006). CITs are constructed as follows: the algorithm tests the relationship between the dependent variable and each explanatory variable and selects the variable with the strongest association. If the association is strong enough (as judged by the significance level α^*), it selects the variable and searches for a value in it, using which the sample is split into two, such that the relationship with the dependent variable is maximized. This procedure of selecting a variable and a split value is repeated in each of the two subsamples until no explanatory variable in any subsample is sufficiently related to the dependent variable. We describe an example of a tree in the figure below.

Figure A: Example Tree



This tree maps out student learning gains for Math using 3 variables: student baseline math score, teacher assessment math score and the percentage of students who say that the teacher reviews concepts at the end of math class. It tells us that if a student's baseline score is less than or equal to 51.67 percent, teachers

reviewing concepts is important (split point at 20%) and gains are higher for those teachers who review concepts more often. On the other hand, for students whose baseline score is greater than 51.67 percent, teacher subject content knowledge (measured by teacher assessment score in math) starts to matter with the split point being 59%. Teachers who score higher in subject content knowledge have higher student learning gains (9.8% vs. 0.69%).

There is a bias versus variance trade-off in decision trees captured by the depth of the tree. Shallower trees will have high bias but low variance in their estimates (due to smoothing) while deeper trees will have low bias (as the tree partitions the sample space into more granular groups) but high variance, as they would be sensitive to small changes in the data. The final depth of a tree is closely linked to the significance level α^*.[9] Irrespective of the value of α^* specified, the tree algorithm still selects the most relevant variable and the most relevant splitting point within that variable, yielding good properties for variable selection. The detailed procedure on how a tree is constructed under CIT is laid out in Annex 2.

Trees generally suffer from two major drawbacks. First, their predictions suffer from high variance as they are sensitive to small changes in the data. Second, any given tree will not select more than a handful of variables during their construction. As a result, many variables (especially in high dimensional settings) do not get a chance to contribute to the tree construction process. To address both these issues, it is best to rely on 'forests' not just 'trees'. Accordingly, we use conditional inference forests (CIF) (Breiman, 2001, Biau and Scornet, 2016).

A forest is simply a collection of many trees (conventionally 100 or 500) and rests on the "wisdom of crowds" logic. When a forest makes a prediction for any given observation, it – loosely speaking - averages the predictions made by each tree. Two tweaks are made when constructing a Forest.

First, only a random sample of predictors are selected when constructing any given tree. This ensures that several trees will not be constructed using similar variables and will therefore not yield correlated predictions. This also allows each explanatory variable to get an adequate chance to prove themselves yielding good properties for variable selection.[10]

Second, a random sample of the training data set is used in the construction of each tree. Due to certain statistical properties that are suitable for stable variable selection we sample the training set without replacement (as discussed in Strobl et al. 2007; 2009, Hothorn et al. 2015).

These two features along with the fact that predictions are averaged across many trees ensures that the estimates of the dependent variable have low variance; and deeper trees can be grown to achieve low bias. Just like trees, certain tuning parameters have to be optimized for when constructing a CIF.[11] This optimization ensures that the CIF performs well out of sample. Annex 2 lays out the detailed procedures and choices for the tuning parameters.

[9] A more complete description of the parameters that finally decide the structure of the tree is provided in Annex 2.
[10] In our analysis, we use the square root of the number of explanatory variables based on convention https://www.stat.berkeley.edu/~breiman/Using_random_forests_v3.00.pdf.
[11] The tuning parameters we optimize for are: (i) the minimum number of observations required to create a split, (ii) the significance level alpha, and (iii) the number of trees that make up the forest.

One drawback with forests is that they cannot be visualized in the same manner as trees. Variable importance measures can however be calculated yielding variables that are most predictive of student learning gains. We measure variable importance using the permutation method described in Strobl et al. (2007).[12] Each explanatory variable is permuted such that its association with the dependent variable is lost.[13] The loss in predictive power caused by permuting a particular variable gives us a measure of its importance when making accurate predictions for student learning gains. We report this measure of variable importance for predicting Math and Kiswahili student learning gains in Tables 4b & 5b, respectively. To make variable importance measures more interpretable, we standardize them such that the most important variable takes a value of one and rank them accordingly. The standardized variable importance numbers shown have been averaged by the number of times a variable emerges as important during 20 model runs (discussed further below). We only choose variables that occur in 14 or more runs to account for any multicollinearity and remove the element of random chance in variable selection (analogous to what Mullainathan and Spiess 2017 do for LASSO).

2.2.2 Least Absolute Shrinkage & Selection Operator (LASSO)

The second supervised machine learning algorithm we use is LASSO (Tibshirani, 1996), perhaps the most well-known to economists. LASSO is a penalized form of regression where the L_1 norm of the coefficient vector β_j is included in the OLS minimization problem[14]:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|$$

The absolute-value penalty term effectively ensures that we are left with only a limited number of non-zero coefficients which means, in effect, that LASSO conducts variable selection. LASSO is computationally feasible on high dimensional data sets and also yields good predictions especially when the true underlying relationship is linear and sparse (Zhao and Yu, 2006; Varian, 2014). The tuning parameter λ decides the number of variables selected by LASSO. A larger lambda implies that more coefficients are set to zero (λ=0 gives us OLS). In our empirical strategy, we choose λ based on k-fold cross validation (details in Annex 2).

The benefit of LASSO is that it provides easily interpretable and sparse models.[15] It also allows us to determine whether the associations between the outcome and the predictors are positive or negative globally (something that tree based methods cannot do). An issue as seen in the practical implementation of LASSO is that it may not select variables in a stable manner when the data set is split differently into the training

[12] As mentioned in Strobl (2007), the permutation method (when used with sampling without replacement) yields unbiased variable importance measures and is also unaffected by different types of predictor variables (categorical or continuous) or their scale of measurement.

[13] After the permutation, the out of bag error rate (MSE^{OOB}) is recalculated for the entire forest. The increase in MSE^{OOB} relative to the baseline out-of-bag-error tells us how important a particular variable was.

[14] We standardize all variables in our data set to have mean = 0 and standard deviation = 1; this is a pre-requisite prior to running LASSO since LASSO may otherwise penalize coefficients of variables measured on a larger scale.

[15] LASSO coefficients should however not be interpreted in the way OLS coefficients are since they are biased towards zero.

and test sets (Mullainathan and Spiess 2017).[16] However, in our case, LASSO conducts variable selection in a stable manner with the most important variables being selected in the vast majority of model runs.

2.2.3 OLS specification

As mentioned above, we benchmark the performance of our two ML algorithms against OLS. The OLS model we use for comparison is as follows:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon_i$$

where,

- Y_i is Average Student Learning Gain for Teacher i
- X_1 is a vector of Student controls such as student age, baseline score, whether a student attends private tuitions in that subject[17] and percentage of students belonging to the lowest asset quartile
- X_2 is a vector of school level covariates such as pupil teacher ratio, school location (urban/rural), annual entitlement grant per school etc.
- X_3 is a vector of school level variables that affect the management and governance of teachers
- X_4 is a vector of controls representing teacher characteristics (who teachers are) such as gender, experience and academic qualifications
- X_5 is a vector representing our variables of interest divided into 3 categories: (i) what teachers do (measured through the TEACH Classroom Observation Tool and Student surveys) (ii) what teachers believe (measured through modules in Teacher surveys representing mental models) and (iii) what teachers know[18] (teacher assessment scores etc.)
- ε_i represents an idiosyncratic error term

2.3 Limitations

Our study suffers from three main limitations. First, we cannot fully address the issue of non-random matching of students, teachers, and schools. A rich set of teacher, student, and school level controls helps, as does the reliance on two cohorts (Grades 2 and 3). Nonetheless, despite our use of a very detailed set of teacher variables, the problem of unobserved variable bias remains. Non-random matching is sufficiently mitigated in models that control for a rich set of covariates (Koedel et al. 2015), and TVA estimates that control for lagged test scores exhibit little to no bias (Chetty et al. 2014a). Second, the total of number of teacher observations in our study (436) is not very high. This is partly because we were unable to match a significant share of teachers in our overall sample of 748 teachers because we could not match them to students due to challenges in data collection and entry. The key challenges in matching teachers and students were those of teacher and student turnover and also significant churn in the students assigned to teachers. Often this churn took place in a way that was not formally documented and was hard to establish credible data on. However, our analysis in Table 1 shows that this attrition of teachers is mostly random on observables. Finally, we rely on ML methods which are appropriate for predictions and associations but

[16] This instability can arise due to multicollinearity or if the underlying true relationship is non-linear.
[17] These are averaged for all the students taught by a given teacher since our unit of observation is the teacher.
[18] A comprehensive list of variables along with their description is available in the Annex 1.

do not establish causality (Athey and Imbens 2016). Accordingly, our interpretation and discussion is mostly around predicting student learning gains from a rich set of teacher, student, and school covariates.

3. Descriptive Analysis of Teachers, Students, and Schools

3.1 School Characteristics

Our school sample is predominantly public (97 percent) and rural (79 percent). The average pupil to-teacher (PTR) ratio is nearly 63 students per teacher. Almost all the public schools (98 percent) received an average capitation grant of TZS 7,154 per pupil (USD 3) in 2019 (also shown in Table 2a).

3.2 Student Characteristics

The average age of students in our sample is around 9 years. About 30 percent of students for any given teacher belong to the lowest quartile of the constructed asset index.[19] On average, students in the sample answered 37 percent of the questions correctly in their Math test and 45 percent in their Kiswahili test. Around 43 percent students were not able to correctly add the numbers 11 and 4. In Kiswahili, about 29 percent students were not able to read the word *paka* (cat) (further details in Table 2b). Average improvement from baseline test score to test score at follow-up was 12 percent for Kiswahili and 19 percent for Math, as shown in Figure 1.

3.3 Teacher Characteristics

The teachers in our sample were selected from the HT-provided teacher roster, based on their subject and grade assignment. Teachers teaching the focal subjects Kiswahili, Math, and English were eligible for sampling; and teaching in Standards 2 and 3 were prioritized.

Who teachers are and what they know

Overall, 56 percent of teachers are female. Around 54 percent of teachers have worked less than 10 years in the teaching profession. The mean years of experience is 12 years. Around 75 percent of teachers report having been trained at the diploma level or lower. On average, teachers in the sample answered 71 percent of the questions correctly in the Kiswahili assessment and 72 percent in the math assessment, as shown in figure 2.

[19] For each student, we take the first principal component score of 8 variables indicating household ownership of the following assets: (i) television, (ii) radio, (iii) electricity, (iv) refrigerator, (v) bed/mattress, (vi) motorbike, (vii) fan, (viii) telephone. We then rank them by quartiles and create a dummy variable which equals 1 if a student belongs to the lowest quartile.

How teachers are managed

The mean reported gross monthly compensation for teachers is TZS 676,184 (USD 293). Education, relative speaking, is one of the better-paid sectors (RISE Baseline Report 2019). However, since 2000, most teachers have faced stagnating purchasing power at best. Also, the salary differences between teacher certification types have increased over this period ((RISE Baseline Report 2019).

Around 36 percent teachers believe that their school regularly recognizes and rewards teacher performance. However, only 25 percent said that student learning outcome is the key metric used by the head teacher to judge their performance. Nearly 84 percent of teachers believe actions are taken in case of poor performance. The action format most often taken, according to teachers, was a warning from the head-teacher. According to teachers, the risk of dismissal or transfer because of poor performance is almost zero.

Around 49 percent agree with the following statement about the school leadership: "*They will recommend me to be transferred or dismissed in case I receive too many bad performance evaluations*." In terms of support received (personal and professional) from the school administration and Government, only 17 percent and 23 percent, respectively, teachers feel satisfied.

What teachers do

For *what teachers do*, we report data from *Teach*. Broadly, teachers in our sample score high (average rating of 3-3.5 out of 5) in providing a supportive learning environment and in setting positive behavioral expectations in the classroom. However, they score low (average rating of 1.5-2 out of 5) on providing students with feedback, in perseverance, and in social and collaborative skills.

We also capture what teachers do through selected questions in the student survey: While 89% of students reported that their teacher explains in another way if they do not understand something, only 61% said that teachers write on their notebooks while correcting their work. About 70% of students said they were afraid of their teacher.

What teachers believe

Finally, for *what teachers believe*, some aspects of teacher beliefs are reflected in the section above, in terms of their perceptions of how they are managed. We also use a dedicated module incorporated within the teacher survey that builds on past cross-country research in this area (Sabarwal and Abu Jawdeh 2018). Using questions from this module, we create six mindset-indices, using principal component analysis.[20]

The notable insights from the mindset modules are summarized here. Around 93 percent of teachers claimed that they could successfully teach all relevant subject content to even the most difficult students. But despite this high self-efficacy, about 40 percent of teachers believed there is little they can do to help a student's learning if they come unprepared from previous grades.

Teachers have nuanced views about test-based accountability; 95 percent of teachers believed that they should receive additional bonuses if their students perform well on exams, but only 50 percent believed that

[20] These are: self-efficacy, locus of control, quality of relationships, positive attitude, reinforcement bias, and support for test-based accountability.

their promotion should depend on their student's performance on exams (for comparisons to other countries see Sabarwal and Abu-Jawdeh 2018).

4. Results

In this section we first lay out the relative outperformance of the two supervised ML algorithms used – CIF and LASSO. Next, we present CIF and LASSO results around which teacher covariates are most predictive of student learning gains in Math and Kiswahili. To analyze the relative importance of ML identified variables further, we present results of Post-CIF and Post-LASSO OLS regressions on the selected set of variables. In other words, we show the results from the OLS regression on the ML-identified parsimonious models.

4.1 Performance of machine learning algorithms

As mentioned in section 2.2.1, we split our data into a training set and a test set (details in Annex 2). We train the models on the Training Set and then use the Mean Squared Error (MSE) from applying the model in the Test Set as the evaluation metric to judge their performance. The MSE is particularly useful for prediction problems like ours because it optimally trades off bias and variance (Kleinberg et al 2015, brief discussion in Annex 2). A lower MSE implies a better prediction out of sample.

For each subject, we divide the MSE^{OLS} by the MSE of our ML methods in order to show the relative MSE. Hence, the relative MSE becomes:

$$Relative\ MSE^{CIF} = \frac{MSE^{OLS}}{MSE^{CIF}}$$

$$Relative\ MSE^{LASSO} = \frac{MSE^{OLS}}{MSE^{LASSO}}$$

A relative MSE greater than one implies that OLS performs poorly out of sample relative to the ML algorithm. This could be either because OLS overfits the data or it makes poor use of the explanatory variables due to issues of high dimensionality. We also derive 95% Confidence Intervals for each model[21] to ensure that our results are not sensitive to the training-test split (see for example Brunori et al. 2018 for another application). Results are presented in Tables 3a and 3b and confidence intervals plots in Figure 3.

We find that both CIF and LASSO outperform OLS in predicting student learning gains out-of-sample in the vast majority of cases. Therefore, they better model the relationship between teacher characteristics and student learning gains vis-à-vis OLS. CIF outperforms OLS by 21 percent for Math and 14 percent for Kiswahili[22] (average outperformance of 18 percent); LASSO outperforms OLS by 23 percent for Math and 21 percent for Kiswahili (average outperformance of 22 percent).

[21] For the purposes of deriving 95% C.I., we run LASSO 200 times and Conditional Inference Forest 100 times.
[22] In very few cases, the lower bound of the confidence interval is lower than 1 suggesting that at times, CIF and LASSO may also be overfitting the data relative to OLS.

We also find that teacher covariates are more effective at predicting student learning gains for Math than for Kiswahili. This can be seen from the absolute value of the Mean Squared Error of the Test Set which is higher for Kiswahili than Math. This is also evident in the overall CIF and LASSO measure of variable importance presented in Tables 4b and 5b. As discussed in Section 2.2, for CIF, we measure variable importance using the permutation method described in Strobl et al. (2007), wherein each explanatory variable is permuted such that its association with the dependent variable is lost. For Math, this standardized score is 0.53 for the variable of highest importance after baseline score (teacher practice of providing written feedback on students work). For Kiswahili this score is 0.19 for the variable of highest importance after baseline score (teacher belief that action is taken against poor teacher performance).

4.2 Key predictors of Student Learning Gains

We report the key predictors for student learning gains using CIF and LASSO in Tables 4a and 5a for Math and Kiswahili, respectively. We also show the detailed results for CIF and LASSO in Tables 4b and 5b, respectively. In these detailed results we show the *standardized variable importance scores* for CIF [23] and the *number of times the variable occurred* (in the 20 runs for CIF and 200 runs for LASSO). For LASSO we also show the sign of the coefficient to show the relationship between student learning gains and a given variable in a global sense.

Finally, we also run a 'Post-CIF' & 'Post-LASSO' (Belloni and Chernozhukov 2013) OLS on only those variables that are selected by our ML models with standard errors clustered at the school level.[24] These results are presented in Tables 4c and 5c for Math and Kiswahili respectively.

There are three broad patterns of note in our results. First, we find that variables of importance for predicting student learning gains are very similar across CIF and LASSO (see Tables 4a and 5a), with the variables of importance identified through LASSO being a subset of those identified by CIF. This is perhaps unsurprising since LASSO, in general, yields sparser models relative to CIF and the latter also tends to select those variables which may be important in interaction with other variables. This alignment between CIF and LASSO, which occurs for both Math and Kiswahili, is noteworthy because the methods approach the prediction and model building process in a very different manner (non-parametric for CIF and parametric for LASSO). Further, LASSO permits a check on whether the variables of importance identified through ML show the expected direction.[25]

Our reported ML results (variables of importance for predicting student learning gains) are also stable across multiple runs. Given that the training set and test set are randomly sampled, it is plausible that variables that are important in one iteration may not be so if the training and test sets are randomly sampled again. Hence, to ensure stability, we run our CIF model 20 times, and only report variables that show up as

[23] For LASSO, given the presence of the L1 Norm in LASSO's minimization problem, its coefficients are downward biased and therefore we do not report absolute coefficients since they cannot be interpreted in the same manner as OLS.

[24] Belloni & Chernozhukov (2013) provide a more technical argument on the statistical properties of the Post-LASSO estimator.

[25] This is not possible in CIF, given that forests are non-parametric in nature and cannot be visualized in the same manner as trees, they do not tell us whether there is a positive or negative relationship between student learning gains and teacher covariates in a global sense. However, the sign of the LASSO coefficients informs us on this.

important in 14 or more runs.[26] Result-stability around variables of importance is more of an issue in LASSO (Mullainathan and Spiess 2017) than in forests with large number of trees (Strobl et al. 2009). Therefore, we ran LASSO 200 times, and present only those variables that show up as important in more than 140 models to remove the role of random chance. This also allows us to account for multicollinearity.

Second, as mentioned in Section 4.1, teacher covariates matter more for predicting student learning gains in Math, compared to Kiswahili. Also, the important predictors are different across the two subjects. It is not surprising that teacher covariates are more important, and differently important, for Math as compared to Kiswahili. Studies have generally found greater variance in teacher effects on achievement in Math than in English (or reading). The difference is due to the large share of language learning that happens at home versus the mostly classroom-based learning that happens for Math (Jackson, Rockoff, and Staiger 2014, Bau and Das 2020).

Third, students' baseline score is by far the most important variable[27] for predicting student learning gains, potentially signaling mean reversion.

4.2.1 Math

What are the most important predictors of student learning gains in Math? After controlling for a student's baseline score, the two most important teacher covariates for predicting student learning gains are: (i) teacher practice of providing written feedback to students on their homework / tests and (ii) teacher belief they can help disadvantaged / struggling students learn.[28] These variables have the strongest importance in CIF and are selected as being important by both CIF and LASSO.

Aside from these top variables, other variables of importance occurring in both CIF and LASSO are as follows. Teachers with training in teaching foundational concepts (Reading, Writing, Counting) have higher student learning gains. In addition to providing written feedback, two other teacher practices are important. First, teachers who ask more open ended questions (the critical thinking construct on *Teach*) have lower student learning gains. This is one of the very few counter-intuitive results we see but is consistent with findings from other *Teach* studies signaling a potential measurement and/or interpretation issue in *Teach* (Filmer et. al 2020).[29] Second, teachers who review concepts taught at the end of class have higher student learning gains.

[26] We run CIF 20 times and LASSO 200 times because CIF, which is a collection of trees, is computationally expensive: 20 model runs for CIF take about 55 minutes to execute while 200 runs of LASSO takes less than 1 minute to execute.

[27] We explicitly include baseline test score as a control variable even though it is used to calculate student learning gains. We do this to account for threshold effects in the underlying distribution. We are therefore allowing for the fact that learning gains might be qualitatively different for a student who increases her test score from 20% to 30% as compared to a student who goes from 70% to 80%.

[28] This can be interpreted as a proxy of whether teachers consider student learning of disadvantaged or struggling students to be within their locus of control. It may also be interpreted as a variable signaling how much ownership teachers take of the learning of struggling or disadvantaged students. It can also be interpreted as a growth mindset indicator for teachers, given that teachers who believe they can improve the learning of struggling/ disadvantaged students may be seen to be having a growth mindset.

[29] For instance, it could be signaling that the practice of asking open-ended questions may produce perverse results if key concepts are not explained well.

In terms of what teachers believe, apart from belief that they can help disadvantaged / struggling students learn, two other factors are important. First, teachers who believe that their career progression and salary be linked to student test performance, have higher student learning gains (teacher belief in test-based accountability). Second, teachers who believe they are the most important stakeholder in assessing progress towards professional targets as compared to other stakeholders yield higher student learning gains (teacher belief in their autonomy).

In addition to these teacher covariates, school location – rural or urban – is also a variable of importance for student learning gains.

The OLS regressions using the variables selected by ML show some interesting insights. In the Post-CIF and Post-LASSO regressions (Table 4b), we find that the teacher belief they can help struggling / disadvantaged students learn is significant at the 1% level. Going from the teacher at the 25th percentile to the 75th percentile on the Locus of Control Index is associated with an increase in student learning gains of 0.21 SD. Moving from the 25th to 75th percentile in the teacher practice of reviewing concepts at the end of the class, is associated with a 0.18 SD gain in student learning gains.

Teacher support for test-based accountability (teacher beliefs) is also significant at the 5% level. Teacher training in teaching foundational concepts (teacher preparation) and teacher practice of asking open-ended questions (critical thinking) are significant at the 5% level. The latter is negatively related to student learning gains, in line with the LASSO result.

4.2.2 Kiswahili

As mentioned above, the degree of influence exerted by teacher covariates on student learning gains for Kiswahili is lower than it is for Math (see Tables 3 and compare Tables 4b & 5b). After controlling for baseline score and student age, the two most important variables for predicting student learning gains are: whether action is taken against poor teacher performance and whether a teacher provides extra help to those students who face difficulties during breaks or after school hours. These are selected by both ML Models.

Another somewhat counter-intuitive result is the negative relationship between teacher beliefs that action is taken against poor performance and student learning gains. Our interpretation, based on contextual information, is that 'action against poor performance' typically refers to action against disciplinary infractions (and not necessarily student learning) and in a vast majority of cases involves a warning from head-teacher. So one way to interpret this finding is that teachers who do not believe 'warnings from head-teachers about disciplinary infractions' is effective action against poor teacher performance are linked with higher student learning gains.

On what teachers know, once again training in teaching foundational concepts (Reading, Writing, Counting) is important. This makes sense because a key focus of 3R Training (the government program aimed at these concepts) is improvement in Kiswahili learning outcomes. On what teachers do, apart from offering extra help, teachers who provide lesson facilitation have higher student learning gains.[30] On what

[30] When teachers clearly articulate lesson objectives, explain content clearly and relate classroom lessons with real life situations.

teachers believe, *the relationships index* which measures teacher beliefs about their relationships with students, colleagues and the head teacher is important.[31]

Analysis of the Post-CIF and Post-LASSO regressions show that in a more parsimonious model, offering extra help to lagging students (teacher practice) and the Relationships index (teacher beliefs) are statistically significant at the 1% level. Moving from the 25th to the 75th percentile in terms of teachers offering extra help to lagging students, is linked to student learning gains of 0.19 SD. Similarly, going from the teacher at the 25th percentile to the 75th percentile on the Relationships Index is associated with an increase in student learning gains of 0.14 SD.

5. Conclusion

Research on teacher effectiveness has struggled to identify observable teacher characteristics that can help explain variation in student performance. In this study, we apply machine learning methods to this problem. Using matched student-teacher data for Grade 2 and 3 students from across 392 schools in Tanzania, we use two ML approaches, Conditional Inference Forest (CIF) and LASSO, to predict student learning gains.

We find that ML approaches outperform the standard OLS model by 14-24 percent in out-of-sample predictions. Further, even though both CIF and LASSO take different model-building approaches, they produce largely consistent results. As expected, student baseline scores are the most predictive of student learning gains, signaling mean reversion in the data.

Our key finding is that specific elements of what teachers know (teacher preparation); what teachers do (teacher practice); and what teachers believe (teacher beliefs) are more strongly predictive of student learning gains than other teacher, student, and school factors, especially in Math. For Math, CIF results show that the teacher practice of providing written feedback on homework/tests and reviewing key concepts at the end of class (measured through student surveys); the teacher belief that they can help disadvantaged and struggling students learn; and teacher preparation around teaching foundational concepts are the most important predictors of student learning gains. However, one counter-intuitive result that merits further investigation is that teachers who score high on fostering critical thinking in classroom observations (for instance by asking more open-ended questions) have lower student learning gains.

For Kiswahili, even though teacher (and other observable) covariates matter less for predicting student learning gains, teacher preparation on teaching foundational skills, teacher practice of providing additional support to struggling students, and teacher belief that they have good relationships within school still emerge as important. Consistent with existing literature, commonly used teacher characteristics such as education, experience, assessment scores etc. do not emerge as important predictors of student learning gains. Outside of teacher variables, rural schools (for both Math and Kiswahili) and older students (for Kiswahili) have stronger student learning gains; but these factors are still less important than the top teacher factors.

[31] Examples of questions include (= 1 if 'Agree'; = 0 if 'Disagree'): (i) I have a good relationship with my students, (ii) I have a good relationship with my colleagues, (iii) I have a good relationship with my Head Teacher.

Our findings show how machine learning can be a powerful tool for addressing some of the hitherto unanswered questions around teacher effectiveness. They may also contribute to the growing interest in understanding and systematically measuring teacher beliefs and behaviors.

No one study can provide a definitive set of guidance, but our results suggest that teacher training programs need to focus more directly on preparing teachers to teach foundational skills, and fostering in them the practice of providing written feedback to students, reviewing key concepts at the end of class, and spending extra time with struggling students. These elements should also be emphasized in teacher supervision and management.

Our findings also demonstrate the importance of systematically measuring and targeting specific aspects of teacher beliefs. Research from education and economics has long shown that teacher beliefs can impact student outcomes directly (Jussim and Harber 2005, Bertrand and Duflo 2017, Sabarwal et. al 2021). However, despite their importance and measurability, there is very little systematic data or discussion on teacher beliefs in the rich literature on education impact evaluations (Sabarwal et. al 2021). This paper demonstrates that teacher beliefs need to be a part of the discussion on improving teacher effectiveness.

Specifically, for effective teaching, it is crucial for teachers to believe that students can in fact learn. Emerging insights from behavioral economics and social psychology demonstrate that these beliefs can be systematically fostered in teachers (and also students themselves). Incorporating these ideas in the design and implementation of teacher programs may help improve teacher effectiveness. However, more research is needed – both on measurement and application - before clear pathways to doing this can be established. Specifically, it is important to understand – what do teachers believe about whether or not disadvantaged students can learn and how best to help them? How malleable are these beliefs? Can they be realistically reshaped to make a big difference for learning outcomes? At which point (pre-service, in-service) would it be best to intervene, if it does make sense to do so?

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*(1), 95-135.

Adelman, M., Haimovich, F., Ham, A., & Vazquez, E. (2018). Predicting school dropout with administrative data: new evidence from Guatemala and Honduras. *Education Economics*, *26*(4), 356-372.

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353-7360.

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*, 685-725.

Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016). Predicting student dropout in higher education. *arXiv preprint arXiv:1606.06364*.

Azam, M. and Kingdon, G.G. (2015). Assessing teacher quality in India. *Journal of Development Economics, 117*, pp.74-83.

Bacher-Hicks, A., Kane, T. J., & Staiger, D. O. (2014). *Validating teacher effect estimates using changes in teacher assignments in Los Angeles* (No. w20657). National Bureau of Economic Research.

Bacher-Hicks, A., Chin, M. J., Kane, T. J., & Staiger, D. O. (2017). *An evaluation of bias in three measures of teacher quality: Value-added, classroom observations, and student surveys* (No. w23478). National Bureau of Economic Research.

Bau, N., & Das, J. (2020). Teacher value added in a low-income country. *American Economic Journal: Economic Policy*, *12*(1), 62-96.

Belloni, A., & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, *19*(2), 521-547.

Bertrand, M., & Duflo, E. (2017). Field Experiments on Discrimination. In *Handbook of Field Experiments*. edited by Banerjee, A. and Duflo, E. Amsterdam, Netherlands: Vol. 1. Elsevier, 309–93.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197-227.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Brunori, P., Hufe, P., & Mahler, D. G. (2018). The roots of inequality: Estimating inequality of opportunity from regression trees. *World Bank Policy Research Working Paper*, (8349).

Buhl-Wiggers, J., Kerwin, J., Smith, J., & Thornton, R. (2017, April). The impact of teacher effectiveness on student learning in Africa. In *Centre for the Study of African Economies Conference*.

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., & Mullainathan, S. (2016). Productivity and selection of human capital with machine learning. *American Economic Review*, *106*(5), 124-27.

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

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American economic review*, *104*(9), 2633-79.

Cruz-Aguayo, Y., Ibarrarán, P., & Schady, N. (2017). Do tests applied to teachers predict their effectiveness?. *Economics Letters*, *159*, 108-111.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., ... & Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27-46.

Dweck, C. S. (2006). *Mindset*. New York: Random House.

Filmer, D., Molina, E., & Wane, W. (2020). *Identifying Effective Teachers: Lessons from Four Classroom Observation Tools*. The World Bank.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The Elements of Statistical Learning* (Vol. 1, No. 10). New York: Springer series in statistics.

Glazerman, S., & Protik, A. (2015). Validating value-added measures of teacher performance. *Association for Public Policy Analysis & Management, November, Miami*.

Gordon, R. J., Kane, T. J., & Staiger, D. (2006). *Identifying effective teachers using performance on the job* (pp. 2006-01). Washington, DC: Brookings Institution.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review*, *100*(2), 267-71.

Heslin, P. A., Latham, G. P., & VandeWalle, D. (2005). The effect of implicit person theory on performance appraisals. *Journal of Applied Psychology*, *90*(5), 842.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, *15*(3), 651-674.

Hothorn, T., & Zeileis, A. (2015). partykit: A modular toolkit for recursive partytioning in R. *The Journal of Machine Learning Research*, *16*(1), 3905-3909.

Jacob, B. A., & Lefgren, L. (2005). *Principals as agents: Subjective performance measurement in education* (No. w11463). National Bureau of Economic Research.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical learning* (Vol. 112, p. 18). New York: springer.

Jussim, L., & K. D. Harber. (2005). Teacher Expectations and Self-Fulfilling Prophecies: Knowns and Unknowns, Resolved and Unresolved Controversies. *Personality and Social Psychology Review, 9* (2): 131–55.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education review*, *27*(6), 615-631.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, *105*(5), 491-95.

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, *47*, 180-195.

Loh, W. Y. (2011). Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, *1*(1), 14-23.

Mbiti, I., Romero, M., & Schipper, Y. (2019). *Designing effective teacher performance pay programs: Experimental evidence from Tanzania* (No. w25903). National Bureau of Economic Research.
McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of educational and behavioral statistics*, *29*(1), 67-101.

Molina, E., Fatima, S. F., Ho, A., Hurtado, C. M., Wilichowksi, T., & Pushparatnam, A. (2018). Measuring Teaching Practices at Scale: Results from the Development and Validation of the Teach Classroom Observation Tool.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87-106.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects?. *Educational evaluation and policy analysis*, *26*(3), 237-257.

Paunesku, D., Walton, G. M., Romero, C., Smith, E. N., Yeager, D. S., & Dweck, C. S. (2015). Mind-set interventions are a scalable treatment for academic underachievement. *Psychological science*, *26*(6), 784-793.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417-458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American economic review*, *94*(2), 247-252.

Rothstein, J. (2014). Revisiting the impacts of teachers. *UC-Berkeley Working Paper*.

Sabarwal, S., & Abu-Jawdeh, M. (2018). What teachers believe: mental models about accountability, absenteeism, and student learning. *World Bank Policy Research Working Paper*, (8454).

Sansone, D. (2019). Beyond early warning indicators: high school dropout and machine learning. *Oxford bulletin of economics and statistics*, *81*(2), 456-485.

Schiltz, F., Masci, C., Agasisti, T., & Horn, D. (2018). Using regression tree ensembles to model interaction effects: a graphical approach. *Applied Economics*, *50*(58), 6341-6354.

Staiger, Douglas O., and Jonah E. Rockoff. 2010. "Searching for effective teachers with imperfect information." *Journal of Economic Perspectives* 24(3): 97-118.)

Strobl, C., Hothorn, T., & Zeileis, A. (2009). Party on!.

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, *8*(1), 1-21.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3-28.

Yeager, D. S., & Dweck, C. S. (2012). Mindsets that promote resilience: When students believe that personal characteristics can be developed. *Educational psychologist*, *47*(4), 302-314.

Zhao, P., & Yu, B. (2006). On model selection consistency of Lasso. *The Journal of Machine Learning Research*, *7*, 2541-2563.

Figure 1: Distribution of Student Learning Gains by Subject

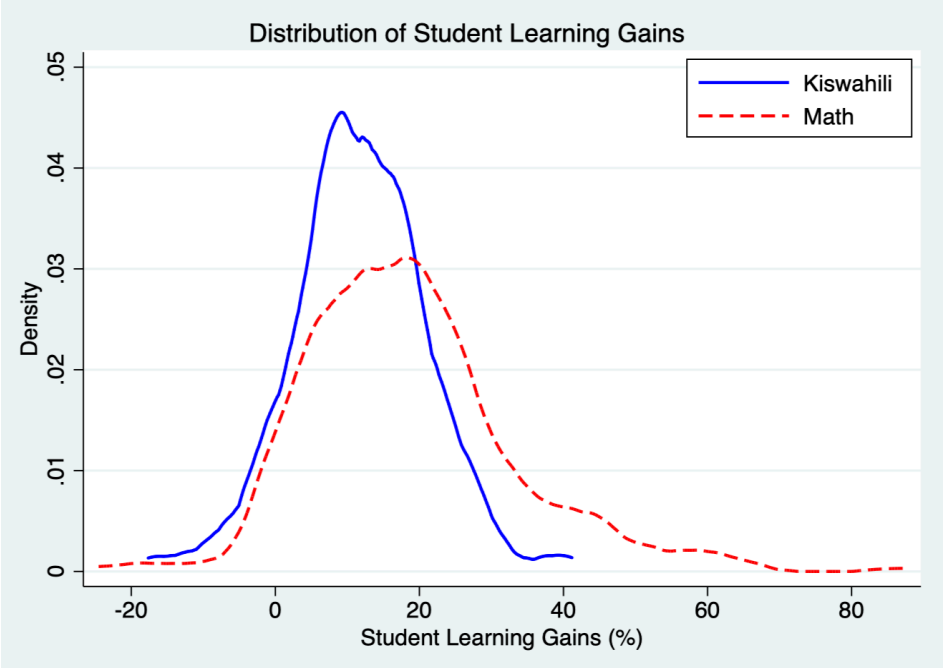


Figure 2: Distribution of Teacher Assessment Scores

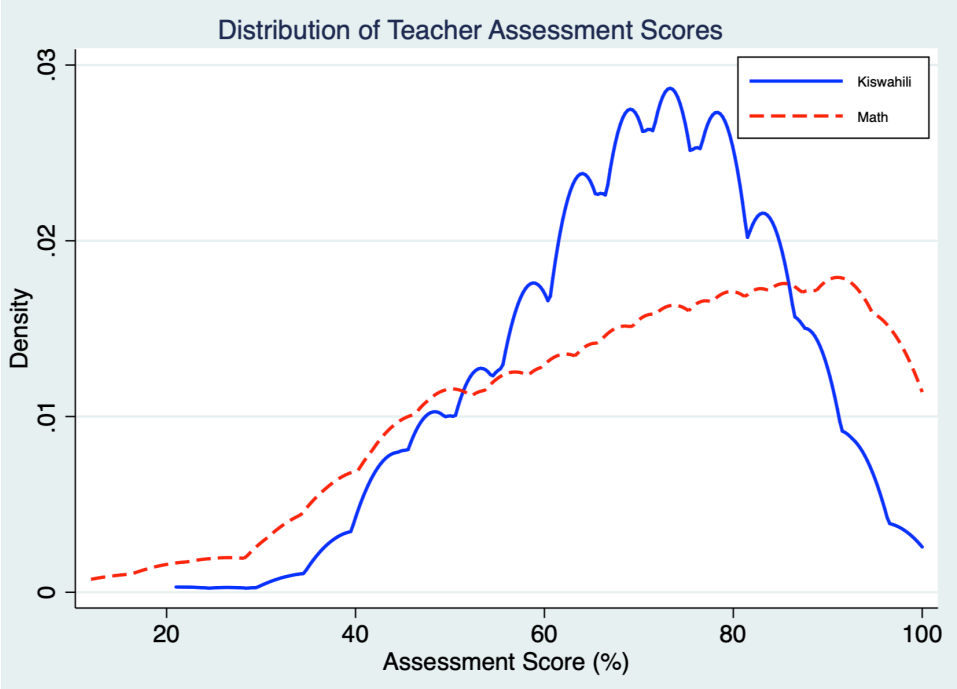


Table 1: Sample characteristics with and without attrition

Variable	Full Sample (N=748)	Sample After Attrition (N=436)	p-value
Teacher gender (= 1 if Male)	0.465	0.440	0.407
What is the highest level of education that you have completed? (=1 if Diploma or Higher)	0.282	0.239	0.097*
Did you specialize in Kiswahili in your teacher training (= 1 if Yes)	0.545	0.550	0.868
Did you specialize in Math in your teacher training (= 1 if Yes)	0.443	0.463	0.489
Have you ever received training in 3Rs (= 1 if Yes)	0.652	0.665	0.656
What is your current gross total compensation per month? (Tanzanian Shillings)	664469.086	644741.355	0.282
Students deserve more of my attention if they: They are lagging behind in classwork (= 1 if Yes)	0.243	0.227	0.524
Teacher's Age	36.916	36.606	0.609
Teacher's Experience	12.496	12.163	0.598
Teacher's Experience in Current School	6.921	6.950	0.941
Teacher Score in Math Assessment (%)	74.552	73.242	0.244
Teacher Score in Kiswahili Assessment (%)	70.306	70.618	0.709

Note: This table presents selected teacher attributes to illustrate balance after attrition. p-values reported in represent the probability of obtaining the corresponding t-test for a null hypothesis that there is no difference in means across the original and current sample. Standard errors are clustered at the school level for this test.

Table 2a: School Level Characteristics (N=306)

Variable	Mean	Standard Deviation
School's Location (= 1 if Urban/Semi-Urban; = 0 if Rural)	0.209	0.407
Pupil Teacher Ratio	62.586	23.427
Amount of Money that is entitled to school as Annual Entitlement Grant? (in Tanzanian Shillings)	7120.916	2989.730
Does the School have a School Management Team? (= 1 if Yes, = 0 if No)	0.778	0.416
Does the School have a Whole School Development Plan?	0.869	0.338

Table 2b: Student Level Characteristics

Variable	Math (N=346)		Kiswahili (N=336)	
	Mean	Standard Deviation	Mean	Standard Deviation
Student Age	8.820	0.907	8.836	0.893
Percentage of students who attend private tuitions in respective subject	4.741	11.260	3.712	10.375
Percentage of students who belong to lowest asset quartile?	30.838	24.437	29.368	24.267
Baseline Test Score (%) in respective subject	36.926	12.111	45.088	25.981

Table 2c: Teacher Level Characteristics

Variable		Math (N=346)		Kiswahili (N=336)	
		Mean	Standard Deviation	Mean	Standard Deviation
Who Teachers Are? (Teacher Characteristics)	Teacher gender (= 1 if Male; = 0 if Female)	0.436	0.497	0.408	0.492
	Teacher's position in the school (= 1 if Head Teacher or Deputy Head Teacher; = 0 if Academic Teacher)	0.156	0.361	0.155	0.360
	Highest Level of Education attained (= 1 if Diploma or Higher, = 0 otherwise)	0.220	0.415	0.244	0.430
	Teacher's Experience (Number of years)	12.777	10.796	12.652	10.870
What Teachers Know? (Teacher Knowledge)	Specialization subject during teacher training? (= 1 if Math/Kiswahili; = 0 for any other subject)	0.480	0.500	0.562	0.497
	Has the teacher received training in 3R? (Reading, Writing, and Counting) (= 1 if Yes; = 0 if No)	0.705	0.457	0.702	0.458
	Is the Teacher provided with information on students' ability at the beginning of the year? (= 1 if Yes; = 0 if No)	0.704	0.447	0.683	0.457
	Does the Teacher have a record of the pupils' continuous assessments? (= 1 if Yes; = 0 if No)	0.327	0.470	0.336	0.473
	Has the Teacher assessed student's curriculum skills using written assessments in the last 5 school days? (= 1 if Yes; = 0 if No)	0.760	0.428	0.747	0.435
	Did the Teacher attend any form of training in the last one year? (= 1 if Yes; = 0 if No)	0.483	0.500	0.509	0.501

	Teacher Assessment Score in Math/Kiswahili (%)	71.465	20.369	69.848	14.059
What Teachers Do? (*Teach* Classroom Observation Tool)	Share of Time teacher spent teaching with medium/high student engagement? (%)	70.809	25.453	72.123	24.771
	Supportive Learning Environment (Range of Values: 1-5)	3.425	0.591	3.414	0.557
	Positive Behavioral Expectations (Range of Values: 1-5)	3.134	0.694	3.128	0.677
	Lesson Facilitation (Range of Values: 1-5)	2.935	0.697	2.961	0.686
	Checks for Understanding (Range of Values: 1-5)	3.116	0.844	3.027	0.822
	Feedback (Range of Values: 1-5)	2.194	0.890	2.149	0.850
	Critical Thinking (Range of Values: 1-5)	2.246	0.718	2.275	0.734
	Autonomy (Range of Values: 1-5)	2.438	0.644	2.443	0.657
	Perseverance (Range of Values: 1-5)	2.048	0.458	2.049	0.467
	Social & Collaborative Skills (Range of Values: 1-5)	1.526	0.738	1.519	0.713
What Teachers Do? (Percentage of students taught by a given teacher who said Yes when asked the following)	If you don't understand something, your teacher explains it another way	88.996	17.340	89.384	15.772
	You are afraid of your teacher	69.578	26.428	69.243	27.216
	At the end of each class, your teacher takes the time to review/discuss	78.734	23.687	77.954	24.131
	When the teacher corrects my work, she writes on my papers to help me	61.977	30.478	62.565	30.933
	Your teacher offers extra help to students who find the subject difficult	79.851	22.886	79.887	23.135
What Teachers Believe?	Students deserve more of my attention if they: They are lagging behind in classwork/homework (1 if Yes; = 0 if No)	0.223	0.417	0.214	0.411

	If students aren't disciplined at home, they aren't likely to accept any discipline at school (1 if Yes; = 0 if No)	0.486	0.501	0.491	.506
	It is okay to be absent as long as I: complete the curriculum OR leave students with work OR doing something useful for the community (= 1 if Yes; = 0 if No)	0.682	0.466	0.679	0.468
	Are any actions taken in case of poor teacher performance? (1 if Yes; = 0 if No)	0.821	0.384	0.857	0.350
	1st PC score of 4 Positive Attitude oriented variables.	0.000	1.176	0.000	1.187
	1st PC score of 5 Incentive oriented variables.	0.000	1.248	-0.000	1.244
	1st PC score of 6 Self-Efficacy oriented variables.	0.000	1.366	0.000	1.370
	1st PC score of 3 Relationship oriented variables.	-0.000	1.055	-0.000	1.038
	1st PC score of 7 Reinforcement tendency variables.	-0.000	1.528	0.000	1.545
	1st PC score of 4 Locus of Control variables.	-0.000	1.066	-0.000	1.081
Teacher Management/School Level Governance	Who is the most important person to assess progress in your professional targets? (= 1 if Teacher says myself; = 0 if Teacher names another person such as Head Teacher)	0.286	0.453	0.277	0.448
	How often does someone from the school leadership observe your classroom? (=1 if 'Once per term'; = 0 if lesser)	0.488	0.501	0.485	0.501
	Does your school regularly recognize or reward teacher performance? (1 if Yes; = 0 if No)	0.361	0.478	0.357	0.477
	What is the one key result HT would assess when rating your job performance?	0.251	0.434	0.256	0.437

	(1 if Exam Results/Learning Progress; = 0 if Other criteria)				
	They will recommend me to be transferred or dismissed in case I receive too many bad performance evaluations? (1 if 'Agree'; = 0 if 'Disagree')	0.500	0.501	0.509	0.501
	Are you satisfied by the support you get from the school administration? (= 1 if Teacher says ‘Satisfied’; = 0 if Teacher says ‘Not Satisfied’)	0.171	0.377	0.176	0.381
	Are you satisfied by the support you get from the Government? (= 1 if Teacher says ‘Satisfied’; = 0 if Teacher says ‘Not Satisfied’)	0.231	0.422	0.223	0.417

Table 3a: Relative Performance of CIF and LASSO vis-à-vis OLS

Test Set	Relative Mean Squared Error	
	Math	Kiswahili
MSE^{OLS}/MSE^{CIF}	1.21 [0.99,1.53]	1.14 [0.85,1.61]
MSE^{OLS}/MSE^{LASSO}	1.23 [0.96,1.63]	1.22 [0.97,1.46]

Note: (i) Figures in parenthesis show 95% C.I. (ii) Relative MSE is defined as $MSE^{OLS}/MSE^{ML\ Model}$

Table 3b: Absolute Performance of CIF and LASSO vis-à-vis OLS

Machine Learning Algorithm	Absolute Mean Squared Error (Test Set)	
	Math	Kiswahili
Conditional Inference Forest	85.30 [54.74,113.14]	175.82 [103.78,292.90]
LASSO	83.38 [55.25,115.43]	161.07 [105.28,224.07]
OLS	102.38 [70.14,140.05]	194.69 [121.98,283.49]

Note: Figures in parenthesis show 95% C.I.

Figure 3: Relative MSE for CIF and LASSO vis-à-vis OLS (95% C.I.)

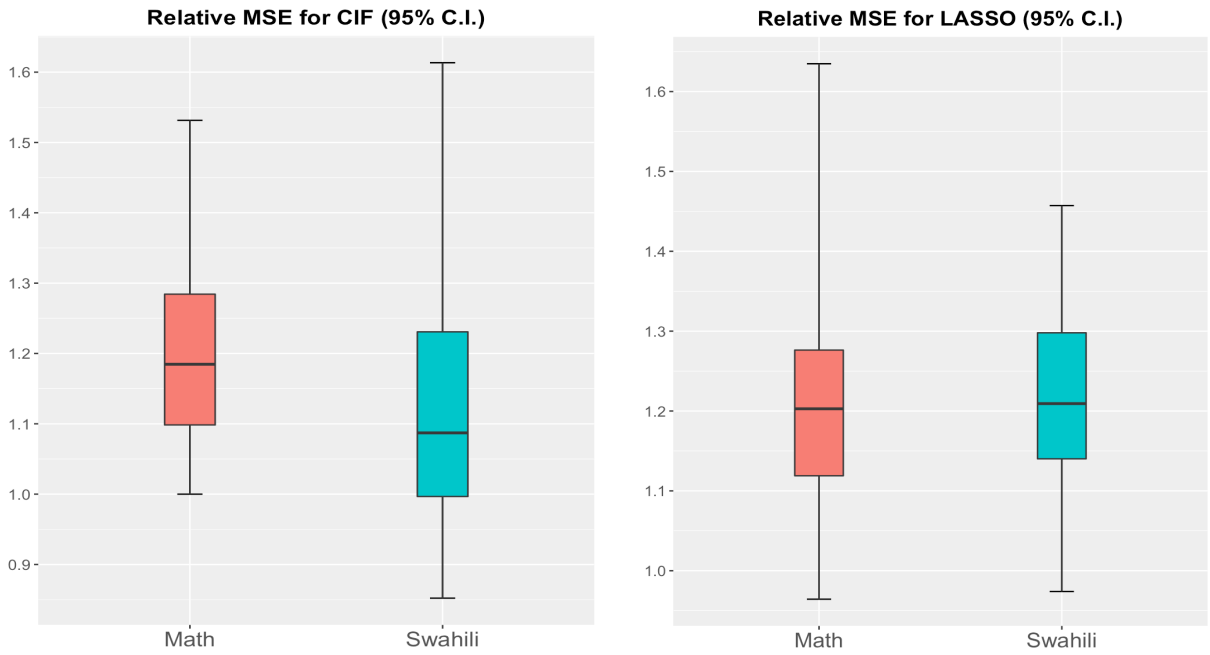


Table 4a: Variable Importance for Math (N=346)

Variable	Conditional Inference Forest	LASSO
1) Baseline Math (%) Score	✓	✓
2) When the math teacher corrects my work, he/she writes on my papers to help me understand (Percentage of students for a given teacher who said Yes when asked this question)	✓	✓
3) Locus of Control Index (Teacher Mental Models) (Teachers who believe they can help disadvantaged / struggling students learn)	✓	✓
4) Critical Thinking (Classroom Observation) (Teacher rated higher if she asks more open ended questions or provides thinking tasks to students)	✓	✓
5) Have you ever received training in 3Rs? (Teachers who received training in the 3R (Reading, Writing, Counting) Program)	✓	✓
6) Teacher Incentive Index (Teacher Mental Models) (Teachers who strongly believe that their career progression and salary is linked to their students' test-score performance)	✓	
7) School in Urban Area	✓	✓
8) At the end of each class, your Math teacher takes the time to review and discuss concepts (Percentage of students for a given teacher who said Yes when asked this question)	✓	✓
9) Who is the most important person to assess your progress towards your professional targets? (= 1 if Teacher says myself; = 0 if Teacher names another person such as Head Teacher)	✓	

Note: (i) Variables in bold are selected as important by both CIF & LASSO (ii) Variables that are important for CIF are those that occur 14 or times in 20 runs of the model; while important variables for LASSO are those that occur more than 140 times in 200 runs of the model.

Table 4b: Detailed Variable Importance for Math – CIF and LASSO (N=346)

Variable	Conditional Inference Forest	LASSO
1) **Baseline Math (%) Score**	0.99 (20)	197 (-)
2) **When the math teacher corrects my work, he/she writes on my papers to help me understand** (Percentage of students for a given teacher who said Yes when asked this question)	0.53 (19)	162 (+)
3) **Locus of Control Index (Teacher Mental Models)** (Teachers who believe they can help disadvantaged / struggling students learn)	0.45 (20)	191 (+)
4) **Critical Thinking (Classroom Observation)** (Teacher rated higher if she asks more open ended questions or provides thinking tasks to students)	0.37 (20)	196 (-)
5) **Have you ever received training in 3Rs** (Teachers who received training in the 3R (Reading, Writing, Counting) Program)	0.25 (18)	184 (+)
6) Teacher Incentive Index (Teacher Mental Models) (Teachers who strongly believe that their career progression and salary be linked to their students' test-performance score higher)	0.23 (19)	
7) **School in Urban Area**	0.20 (18)	147 (-)
8) **At the end of each class, your Math teacher takes the time to review and discuss concepts** (Percentage of students for a given teacher who said Yes when asked this question)	0.20 (16)	161 (+)
9) Who is the most important person to assess your progress towards your professional targets? (= 1 if Teacher says myself; = 0 if Teacher names another person such as Head Teacher)	0.14 (17)	

Note: (i) Variables in bold are selected as important by both CIF & LASSO (ii) Numbers for CIF show relative variable importance (variables ranked in terms of loss in predictive power if a given variable is permuted) and figures in parenthesis show the number of times the variable showed up in 20 runs of the model (iii) Figures for LASSO show the number of times the variable appeared in 200 runs of the model along with the coefficient sign in parenthesis

Table 4c: OLS Regressions of Student Learning Gains for Math on variables selected by CIF & LASSO (Post-CIF & Post-LASSO) (N=346)

Variable	Variables selected by CIF (A)	Variables selected by CIF and LASSO (B)	Variables selected by either CIF or LASSO (AUB)
1) **Baseline Math Score**	-0.194***	-0.186***	-0.194***
	(0.0564)	(0.0564)	(0.0564)
2) **When the math teacher corrects my work, he/she writes on my papers to help me understand**	0.0663	0.0704	0.0663
	(0.0542)	(0.0549)	(0.0542)
3) **Locus of Control Index (Teacher Mental Models)**	0.151***	0.146***	0.151***
	(0.0527)	(0.0536)	(0.0527)
4) **Critical Thinking (Classroom Observation)**	-0.125**	-0.127**	-0.125**
	(0.0523)	(0.0527)	(0.0523)
5) **Have you ever received training in 3Rs**	0.128***	0.120**	0.128***
	(0.0494)	(0.0497)	(0.0494)
6) Teacher Incentive Index (Teacher Mental Models)	0.100**		0.100**
	(0.0499)		(0.0499)
7) **School in Urban Area**	-0.0879*	-0.0872	-0.0879*
	(0.0530)	(0.0540)	(0.0530)
8) **At the end of each class, your Math teacher takes the time to review and discuss concepts**	0.126**	0.123**	0.126**
	(0.0562)	(0.0567)	(0.0562)
9) Who is the most important person to assess your progress towards your professional targets?	0.0895*		0.0895*
	(0.0494)		(0.0494)
R-squared	0.144	0.127	0.144

Note: We report robust standard errors clustered at the School level in parenthesis | Variables in Bold have been selected as important by both CIF & LASSO | Variables have been standardized with Mean = 0 and Std. Dev. = 1 | ***, **, and * indicate significance at the 1, 5, and 10 critical level, respectively |

Table 5a: Variable Importance for Kiswahili (N=336)

Variable	Conditional Inference Forest	LASSO
1) **Baseline Kiswahili Score**	✓	✓
2) **Are any actions taken in case of poor teacher performance?** (= 1 if Teacher says Yes; = 0 if Teacher says No)	✓	✓
3) Student Age	✓	
4) Have you received training in 3Rs? (Teachers who received training in the 3R (Reading, Writing, Counting) Program)	✓	
5) **Your Kiswahili teacher offers extra help to students who find the subject difficult** (Percentage of students for a given teacher who said Yes when asked this question)	✓	✓
6) Are you satisfied by the support you get from the school administration? (= 1 if Teacher says 'Satisfied'; = 0 if Teacher says 'Not Satisfied')	✓	
7) It is okay to be absent as long as I: complete the curriculum OR leave students with work OR doing something useful for the community (= 1 if Teacher answers 'Yes'; = 0 if Teacher answers 'No')	✓	
8) Lesson Facilitation (Classroom Observation) (Teacher rated higher if lesson objectives are clearly articulated, explanation of content is clear, teacher connects lessons to real life)	✓	
9) Pupil Teacher Ratio	✓	
10) Relationships Index (Teachers who strongly believe that they have a good relationship with their students, colleagues & head teacher score higher)	✓	✓
11) Are you satisfied by the support you get from the Government? (= 1 if Teacher says 'Satisfied'; = 0 if Teacher says 'Not Satisfied')	✓	

Note: (i) Variables in bold are selected as important by both CIF & LASSO (ii) Variables that are important for CIF are those that occur 14 or times in 20 runs of the model; while important variables for LASSO are those that occur more than 140 times in 200 runs of the model

Table 5b: Detailed Variable Importance for Kiswahili – CIF and LASSO

Variable	Conditional Inference Forest	LASSO
1) **Baseline Kiswahili Score**	1 (20)	200 (-)
2) **Are any actions taken in case of poor teacher performance?** (= 1 if Teacher says Yes; = 0 if Teacher says No)	0.19 (20)	178 (-)
3) Student Age	0.15 (20)	
4) Have you received training in 3Rs? (Teachers who received training in the 3R (Reading, Writing, Counting) Program)	0.06 (18)	
5) **Your Kiswahili teacher offers extra help to students who find the subject difficult** (Percentage of students for a given teacher who said Yes when asked this question)	0.05 (20)	188 (+)
6) Are you satisfied by the support you get from the school administration? (= 1 if Teacher says 'Satisfied'; = 0 if Teacher says 'Not Satisfied')	0.05 (18)	
7) It is okay to be absent as long as I: complete the curriculum OR leave students with work OR doing something useful for the community (= 1 if Teacher answers 'Yes'; = 0 if Teacher answers 'No')	0.04 (16)	
8) Lesson Facilitation (Classroom Observation) (Teacher rated higher if lesson objectives are clearly articulated, explanation of content is clear, teacher connects lessons to real life)	0.04 (14)	
9) Pupil Teacher Ratio	0.02 (19)	
10) **Relationships Index** (Teachers who strongly believe that they have a good relationship with their students, colleagues & head teacher score higher)	0.02 (14)	145 (+)
11) Are you satisfied by the support you get from the Government? (= 1 if Teacher says 'Satisfied'; = 0 if Teacher says 'Not Satisfied')		144 (+)

Note: (i) Variables in bold are selected as important by both CIF & LASSO (ii) Numbers for CIF show relative variable importance (variables ranked in terms of loss in predictive power if that variable is permuted) and figures in parenthesis show the number of times the variable showed up in 20 runs of the model (iii) Figures for LASSO show the number of times the variable appeared in 200 runs of the model along with the coefficient sign in parenthesis

Table 5c: OLS Regressions of Student Learning Gains for Kiswahili on variables selected by CIF & LASSO (Post-CIF & Post-LASSO; N=336)

Variable	Variables selected by CIF (A)	Variables selected by CIF and LASSO (B)	Variables selected by either CIF or LASSO (AUB)
1) **Baseline Kiswahili Score**	-0.491*** (0.0624)	-0.491*** (0.0541)	-0.489*** (0.0621)
2) Student Age	-0.0303 (0.0450)		-0.0283 (0.0447)
3) **Are any actions taken in case of poor teacher performance?**	-0.115** (0.0539)	-0.120** (0.0518)	-0.107* (0.0545)
4) Have you received training in 3Rs?	0.0154 (0.0471)		0.0205 (0.0465)
5) **Your Kiswahili teacher offers extra help to students who find the subject difficult**	0.139*** (0.0471)	0.138*** (0.0458)	0.130*** (0.0475)
6) Are you satisfied by the support you get from the school administration? (= 1 if Teacher says 'Satisfied'; = 0 if Teacher says 'Not Satisfied')	0.1875 (0.134)		0.151 (0.133)
7) It is okay to be absent as long as I: complete the curriculum OR leave students with work OR doing something useful for the community	0.0657 (0.0448)		0.0649 (0.0448)
8) Lesson Facilitation (Classroom Observation)	0.0727 (0.0454)		0.0746* (0.0446)
9) Pupil Teacher Ratio	-0.0492 (0.0499)		-0.0478 (0.0492)
10) **Relationships Index**	0.114*** (0.0381)	0.104*** (0.0374)	0.115*** (0.0382)
11) Are you satisfied by the support you get from the Government?		0.0805 (0.0514)	0.0813 (0.0511)
R-squared	0.315	0.307	0.321

Note: We report robust standard errors clustered at the School level in parenthesis | Variables in Bold have been selected as important by both CIF & LASSO | Variables have been standardized with Mean = 0 and Std. Dev. = 1 | ***, **, and * indicate significance at the 1, 5, and 10 critical level, respectively

Annex 1: Variable Description

Variable Category	Variable Description
Student Characteristics	Average Student Age for a given Teacher
	Percentage of students taught by a given teacher who attend private tuitions in Math/Kiswahili
	Percentage of students taught by a given teacher who belong to lowest asset quartile? (based on quartile ranking of first Principal Component Score of dummy variables indicating ownership of several assets such as TV, Land, Electricity etc.)
	Baseline Math/Kiswahili Test Score (%)
School Characteristics	Pupil Teacher Ratio
	Does the School have a School Management Team? (= 1 if Yes, = 0 if No)
	Does the School have a Whole School Development Plan?
	Amount of Money that is entitled to school as Annual Entitlement Grant? (in Tanzanian Shillings)
	School's Location (= 1 if Urban/Semi-Urban; = 0 if Rural)
Who Teachers Are? (Teacher Characteristics)	Teacher gender (= 1 if Male; = 0 if Female)
	Teacher's position in the school (= 1 if Head Teacher or Deputy Head Teacher; = 0 if Academic Teacher)
	Highest Level of Education attained (= 1 if Diploma or Higher, = 0 otherwise)
	Teacher's Experience (Number of years)
What Teachers Know? (Teacher Knowledge)	Specialization subject during teacher training? (= 1 if Math/Kiswahili; = 0 for any other subject)
	Has the teacher received training in 3R? (Reading, Writing, and Counting) (= 1 if Yes; = 0 if No)
	Teacher Assessment Score in Math/Kiswahili (%)
	Did the Teacher attend any form of training in the last one year? (= 1 if Yes; = 0 if No)

	Is the Teacher provided with information on students' ability at the beginning of the year? (= 1 if Yes; = 0 if No)
	Does the Teacher have a record of the pupils' continuous assessments? (= 1 if Yes; = 0 if No)
	Has the Teacher assessed student's curriculum skills using written assessments? (= 1 if Yes; = 0 if No)
What Teachers Do? (*Teach* Classroom Observation Tool)	Supportive Learning Environment (Range of Values: 1-5)
	Positive Behavioral Expectations (Range of Values: 1-5)
	Lesson Facilitation (Range of Values: 1-5)
	Checks for Understanding (Range of Values: 1-5)
	Feedback (Range of Values: 1-5)
	Critical Thinking (Range of Values: 1-5)
	Autonomy (Range of Values: 1-5)
	Perseverance (Range of Values: 1-5)
	Social & Collaborative Skills (Range of Values: 1-5)
	Share of Time teacher spent teaching with medium/high student engagement? (%)
What Teachers Do? (Percentage of students taught by a given teacher who said Yes when asked the following)	If you don't understand something, your Math/Kiswahili teacher explains it another way
	You are afraid of your Math/Kiswahili teacher
	At the end of each class, your Math/Kiswahili teacher takes the time to review/discuss
	When the Math/Kiswahili teacher corrects my work, she writes on my papers to help me
	Your Math/Kiswahili teacher offers extra help to students who find the subject difficult
	Positive Attitude Index 1st Principal Component score of four positive attitude variables:

What Teacher's Believe?	(= 1 if Agree; = 0 if Disagree) 1) I am fully satisfied with my current job. 2) My students' learning/achievement motivates me to carry on teaching. 3) My workload is manageable 4) If I could start over I would choose teaching as a career?
	Teacher Incentive Index 1st Principal Component score of 5 incentive oriented variables: (= 1 if Agree; = 0 if Disagree) 1) If my students perform well on official external exams, I should receive an additional bonus 2) My promotion should partly be dependent on my student's performance on tests. 3) The main factor used to assess my performance as a teacher should be my students (= 1 if bonuses; = 0 for capitation grants) 4) Would you prefer teacher bonuses or school level capitation grants? (= 1 for bonus increase; = 0 for flat increase) 5) Would you prefer a flat increase in salaries of all teachers or a bonus component for performance?
	Self Efficacy Index 1st Principal Component score of 6 self-efficacy oriented variables: (= 1 if Agree; = 0 if Disagree) 1) I can successfully teach all relevant subject content to even the most difficult students 2) I can find creative ways to cope with difficulties such as budget cuts 3) I try new ways of teaching in class. 4) Through my teaching I can help students overcome their constraints/difficulties. 5) I can maintain a positive relationship with parents even when tensions arise. 6) I am convinced that I can help address my students' needs.
	Relationships Index 1st Principal Component score of 3 Relationship oriented variables: (= 1 if 'Agree'; = 0 if 'Disagree') 1) I have a good relationship with my students. 2) I have a good relationship with my colleagues. 3) I have a good relationship with my Head Teacher.
	Reinforcement Bias Index 1st Principal Component score of 7 Reinforcement biased tendency variables: Students deserve more of my attention if they (1 if Yes; = 0 if No): 1) Are motivated to learn 2) Attend school regularly 3) Come to school with the necessary material 4) Have the necessary concepts and foundations from previous classes 5) Their parents are involved in the education of their child

	6) Their parents are willing to invest the necessary financial resources in their child's education 7) They are performing well in my class
	Locus of Control Index 1st Principal Component score of 4 Locus of Control oriented variables: There is little I can do to help a student's learning if (= 1 if No; = 0 if Yes): 1) Students come unprepared from previous grades 2) Parents do not seek feedback from the teacher on student performance 3) Parents do not have the necessary education to help their child be more successful at school (1 = 'Disagree'; 0 = 'Agree') 4) If parents would do more for their children, I could do more
	It is okay to be absent as long as I: complete the curriculum OR leave students with work OR doing something useful for the community (= 1 if Yes; = 0 if No)
	Students deserve more of my attention if they: They are lagging behind in classwork/homework (1 if Yes; = 0 if No)
	Are any actions taken in case of poor teacher performance? (1 if Yes; = 0 if No)
	If students aren't disciplined at home, they aren't likely to accept any discipline at school (1 if Yes; = 0 if No)
Teacher Management/School Level Governance	Who is the most important person to assess progress in your professional targets? (= 1 if Teacher says myself; = 0 if Teacher names another person such as Head Teacher)
	How often does someone from the school leadership observe your classroom? (=1 if 'Once per term'; = 0 if lesser)
	Does your school regularly recognize or reward teacher performance? (1 if Yes; = 0 if No)
	What is the one key result HT would assess when rating your job performance? (1 if Exam Results/Learning Progress; = 0 if Other criteria)
	They will recommend me to be transferred or dismissed in case I receive too many bad performance evaluations? (1 if 'Agree'; = 0 if 'Disagree')
	Are you satisfied by the support you get from the school administration? (= 1 if Teacher says 'Satisfied'; = 0 if Teacher says 'Not Satisfied')
	Are you satisfied by the support you get from the Government? (= 1 if Teacher says 'Satisfied'; = 0 if Teacher says 'Not Satisfied')

Annex 2: Details and Notes on Methodology

A.2.1 Assessing Model Performance

In order to assess model performance, we follow the commonplace practice in machine learning of splitting the data set into a training set with and a test set. We then calculate the Mean Squared Error of the Test Set to assess model performance. The detailed procedure is as follows:

1. Use the sample to create two non-overlapping randomly sampled data sets: a training set with 80% observations where $i^{-T} \in \{1,\ldots,N^{-T}\}$; $N^{-T} = 4/5N$ and the remaining 20% to form the test set with $i^{T} \in \{1,\ldots,N^{T}\}$; $N^{T} = 1/5N$
2. Run the three models: CIF, LASSO and OLS on the training set. This yields a prediction function that characterizes how the explanatory variables are associated with the outcome variable $\widehat{f}\,(X^{-T})$
3. Pass the test set values of the explanatory variables into the prediction function created in step 2 to yield estimates/predictions of the outcome variable $\hat{y}^{test} = \widehat{f}\,(X^{T})$
4. Calculate the Mean Squared Error of the test sample

$$MSE^{Test} = \frac{1}{N^{T}} \sum_{i \in T} [y_i - \hat{y}^{test}]^2$$

A 2.2 Conditional Inference Trees & Forests

Decision Trees partition the sample into M mutually exclusive groups through recursive binary splitting. Based on a splitting criterion, they continue to partition the sample into two until a pre-defined criterion is no more fulfilled (such as an information gain or a minimum improvement in RSS) or until a pre-specified threshold is reached (such as a minimum number of sample required to create further splits). Once every observation is part of a group that is common in the expression of the covariate space $\mathbf{X}$: $(X_1, X_2, \ldots X_k)$, for a given vector of the dependent variable $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$, we get the vector of predicted values $\hat{\boldsymbol{y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$, where

$$\hat{y}_i = \frac{1}{N_m} \sum_{j=1}^{m} y_j \quad y \in G_m$$

Conditional Inference Algorithm

1. Choose a significance level α*
2. Test the null hypothesis for independence of the density function: $H_0^{x^p}: D(Y|x^p) = D(Y)$ for all $x^p \in \mathbf{X}$ and obtain a *p*-value associated with each test, p^{x^p}
    a. Adjust the *p*-values for multiple hypothesis testing using the Bonferroni correction
3. Select the variable x^* with the lowest *p*-value
    a. If $p^{x^*} > \alpha^*$: stop the tree making process
    b. If $p^{x^*} < \alpha^*$: continue by selecting x^* as the splitting variable
4. Test the null hypothesis for independence of the density function between the sub-samples for each possible binary splitting point *s* amongst the values taken by x^* and obtain a *p*-value corresponding to each splitting point
    a. Split the sample based on x^* by selecting s^*, the splitting point having the lowest *p*-value
5. Repeat steps 2-4 for each of the resulting sub-samples until
    a. The dependent variable is independent of each explanatory variable in every sub-sample OR
    b. The number of observations left to create further splits is lower than a pre-specified threshold (usually 5 or 10) (Minimum number of samples required to make a split). This ensures smoothness in predictions made by the tree, OR
    c. The tree reaches a pre-specified maximum depth (the longest paths between the root and a leaf)

The final structure and depth of a tree is dependent on 3 hyper-parameters – the minimum number of observations required to create a split, the maximum depth of the tree and the significance level alpha. The first two are related to each other: trees assigned to have shallower depths will automatically have large number of observations in each split and conversely, if the minimum number of observations required to create a split is kept large then that would automatically result in shallower trees.

The variant of the decision tree used in this study and highlighted above is Conditional Inference Trees. Unlike standard decision trees, conditional trees select those variables and splits that are most related to the dependent variable in a statistically significant sense. Moreover, as pointed out by Hothorn et al. (2006), standard decision trees and random forests are biased towards selecting variables that offer more splitting points: in a data set consisting of continuous and dummy variables, they would tend to select more continuous variables during tree construction. Therefore, for the purposes of our study, we use Conditional Inference Trees and Forests as they would likely yield us with variables that best predict student learning gains.

Just like trees, we need to specify tuning parameters when constructing a conditional forest: (i) the minimum number of observations required to create a split (analogous to the maximum depth of a tree) (iii) the significance level alpha & (iv) the number of trees that make up the forest. Optimal selection of tuning parameters ensures that the model does not overfit the data and performs well out-of-sample relative to in-sample performance. Since our goal is to build the best predictive model for student learning gains that performs well out of sample, we select the values of the tuning parameters that minimize the Mean Squared Error of the Out-of-Bag (OOB) sample (MSE^{OOB}).

Samples of the training set that are not used in the construction of a given tree (due to sampling without replacement) are known as Out-of-Bag samples (OOB). Predictions are then made for these OOB samples using only those trees that did not contain them. The mean squared error of such predictions are then averaged at the forest level across every OOB sample to yield us MSE^{OOB}. We use MSE^{OOB} for parameter tuning (as described above) and for variable importance through the permutation method (Strobl et al. 2007) (described in footnote 16).

A.2.3 Optimal tuning parameters and a discussion on Mean Squared Error

Selecting optimal tuning parameters for Conditional Inference Forests

In this section, we highlight our selection of the optimal tuning parameters for Conditional Inference Forests (α^*, B^*, s^*) for Math & Kiswahili where α represents the significance level, B is the number of trees used to build the Forest and s is the minimum sample size required in order to create splits (minimum split criteria).

In order to select (α^*, B^*, s^*) in a data-driven manner, we follow these steps:

1. Create a grid of values for each tuning parameter
    a. α = [0,0.20,0.40,0.60,0.80,0.90,0.95,0.99]
    b. B = [500,750,1000,1250,1500]
    c. S = [5,10,20]
2. Run the Conditional Inference Forest on each combination of values of the 3 tuning parameters and calculate the Out-of-Bag Mean Squared Error (MSE^{OOB}) for each such Forest.
3. Select the set of tuning parameters whose Conditional Inference Forest has the lowest MSE^{OOB}

From our analysis, we find that for Math, the optimal tuning parameters are: $\{\alpha^*= 0, B^*= 1000, s^* = 20\}$ and for Kiswahili, the optimal tuning parameters are $\{\alpha^*= 0, B^*= 1250, s^* = 10\}$. We therefore show relative outperformance (vis-à-vis OLS) and variable importance measures for Math & Kiswahili by constructing the Conditional Inference Forest built using these tuning parameters.

Selecting optimal tuning parameters for LASSO

Given that the LASSO minimization problem includes a penalty term in the form of the L_1 norm of the coefficient vector β_j, the tuning parameter λ plays a key role in variable selection. Hence in order to select the optimal λ we conduct k-fold cross validation (k=10) using the entire data set (rather than splitting the data between a training and test set). Under some minimal assumptions, k-fold cross validation provides unbiased estimates of the out-of-sample MSE (Friedman et al. 2009). The procedure is laid out as follows:

1. We randomly split the the set of observations into k groups or folds of approximately equal size.
2. The first fold is treated as a test set, and LASSO is implemented on the remaining $k - 1$ folds.
3. The mean squared error of the test set fold, MSE_1, is then calculated

4. This procedure is repeated k times such that each of the k folds created becomes a test set. This process results in k estimates of MSE, MSE_1, MSE_2, . . . , MSE_k. The k-fold Cross Validation estimate is computed by averaging MSE_1, MSE_2, . . . , MSE_k.
5. We repeat Steps 1-4 for each value of λ and select the λ which has the minimum Mean Squared Error estimate. We follow the approach of specifying potential λ values as given in Hastie et al. (2013) and consider 100 values in the range of $[0.0001, 10^5]$

From our analysis, we find that for Math, the optimal $\lambda = 0.53$ and for Kiswahili, the optimal $\lambda = 1.23$. In figures A.1 and A.2 we plot the k-fold cross validation Mean Squared Error against the log(λ) values. The numbers at the top tell us the number of variables chosen by the LASSO model as being important.

Figure A.1: Optimal Lambda Plot for Math

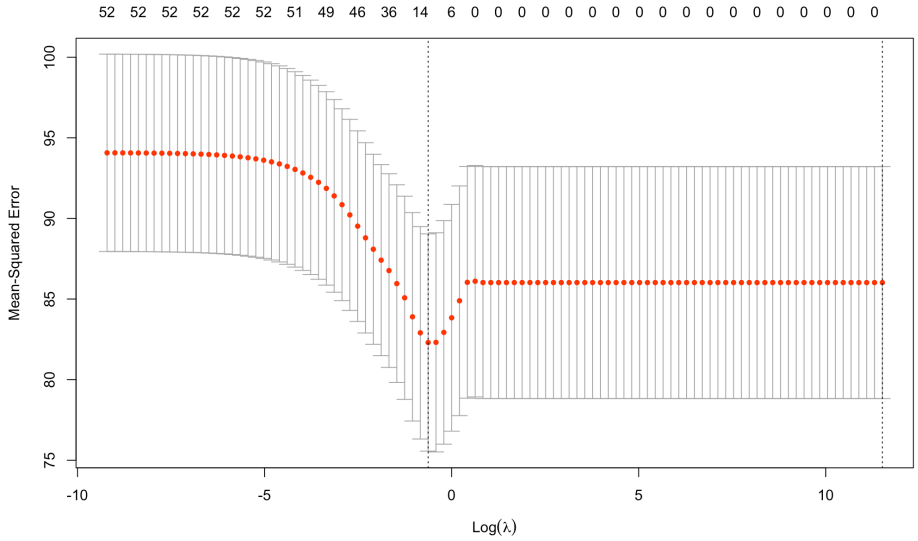
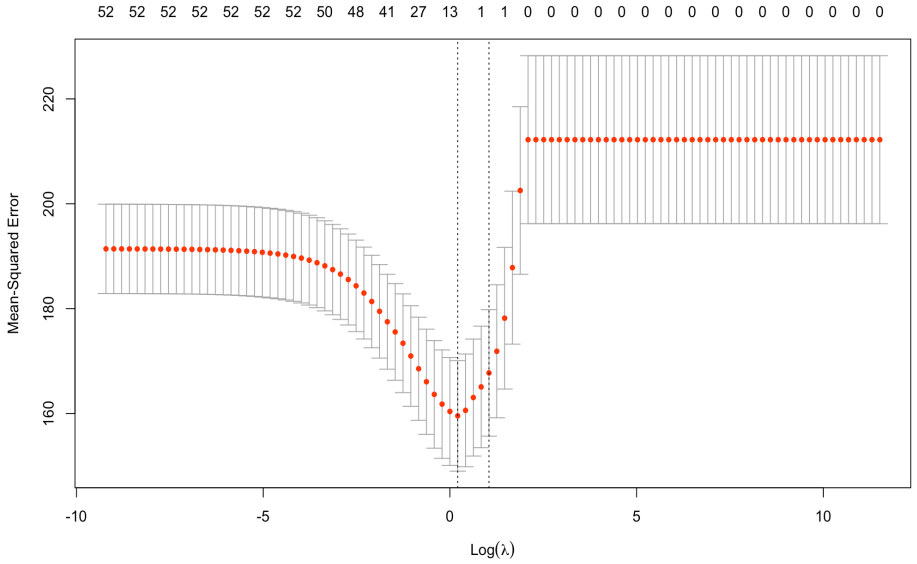


Figure A.2: Optimal Lambda Plot for Kiswahili



Discussion on Mean Squared Error

For any given prediction problem, we are interested in how well our model predicts data that it has never seen before. In order to evaluate how accurately a given Machine Learning model predicts the dependent variable out of sample, we need to account not just for the bias in its prediction (how close is the true value to the predicted value) but also for the variance in its predictions (the sensitivity of our predictions to changes in the underlying data) since any out of sample data is part of the true population.

The Mean Squared Error can be written as:

$$E[(\hat{f}(x) - y)^2]$$

Let's consider the Mean Squared Error at a new data point (x_0,y_0) which can be further decomposed into 3 terms

$$E[(\hat{f}(x) - E[\hat{y}_0])^2] + (E[\hat{y}_0] - y)^2 + Var(\varepsilon)$$

The first term represents Variance which tells us the amount by which the model's prediction would change if we estimated it on a

different data set, the second term represents Bias squared and third term represents the irreducible error term. Given that the first two are non-negative, the MSE can never lie below Var(ε). In order to yield accurate predictions, a machine learning model should minimize the expected mean squared error of the test set. Therefore, it should have *low variance* and *low bias* in its predictions. In general, more flexible models will have lower bias but higher variance whereas less-flexible models like OLS will have low bias but high variance. This trade-off between bias and variance is recurrent in Machine Learning. In fact, since OLS is the best linear unbiased estimator, it does not allow for any trade-off as it sets the second term to zero.

Annex 3: Modeling TVA (traditional approach)

Step 1: Regress follow-up scores on baseline scores along with student-level controls and get teacher Fixed Effects (TVA) (student-level regression). Robust standard errors clustered at school level.
Step 2: Run the teacher fixed effects (TVA) through the Machine Learning algorithms (CIF & LASSO) with teacher-level and school-level covariates.

Math (N=346)

Variable	Selected by CIF or LASSO (current model)	Selected by CIF or LASSO (traditional TVA model)
1) Baseline Math (%) Score	✓	Not applicable (used in calculating TVA)
2) When the math teacher corrects my work, he/she writes on my papers to help me understand (Percentage of students for a given teacher who said Yes when asked this question)	✓	✓
3) Locus of Control Index (Teacher Mental Models) (Teachers who believe they can help disadvantaged / struggling students learn)	✓	✓
4) Critical Thinking (Classroom Observation) (Teacher rated higher if she asks more open ended questions or provides thinking tasks to students)	✓	✓
5) Have you ever received training in 3Rs? (Teachers who received training in the 3R (Reading, Writing, Counting) Program)	✓	
6) Teacher Incentive Index (Teacher Mental Models) (Teachers who strongly believe that their career progression and salary is linked to their students' test-score performance)	✓	✓

7) School in Urban Area	✓	Not applicable (school level variables interpreted as controls)
8) At the end of each class, your Math teacher takes the time to review and discuss concepts (Percentage of students for a given teacher who said Yes when asked this question)	✓	✓
9) Who is the most important person to assess your progress towards your professional targets? (= 1 if Teacher says myself; = 0 if Teacher names another person such as Head Teacher)	✓	

Kiswahili (N=336)

Variable	Selected by CIF or LASSO (current model)	Selected by CIF or LASSO (traditional TVA model)
1) Baseline Kiswahili Score	✓	Not applicable (used in calculating TVA)
2) Are any actions taken in case of poor teacher performance? (= 1 if Teacher says Yes; = 0 if Teacher says No)	✓	✓
3) Student Age	✓	Not applicable (used in calculating TVA)
4) Have you received training in 3Rs? (Teachers who received training in the 3R (Reading, Writing, Counting) Program)	✓	✓
5) Your Kiswahili teacher offers extra help to students who find the subject difficult (Percentage of students for a given teacher who said Yes when asked this question)	✓	✓
6) Are you satisfied by the support you get from the school administration? (= 1 if Teacher says 'Satisfied'; = 0 if Teacher says 'Not Satisfied')	✓	✓
7) It is okay to be absent as long as I: complete the curriculum OR leave students with work OR doing something useful for the community (= 1 if Teacher answers 'Yes'; = 0 if Teacher answers 'No')	✓	✓
8) Lesson Facilitation (Classroom Observation)	✓	✓

(Teacher rated higher if lesson objectives are clearly articulated, explanation of content is clear, teacher connects lessons to real life)		
9) Pupil Teacher Ratio	✓	✓
10) Relationships Index (Teachers who strongly believe that they have a good relationship with their students, colleagues & head teacher score higher)	✓	
11) Are you satisfied by the support you get from the Government? (= 1 if Teacher says 'Satisfied'; = 0 if Teacher says 'Not Satisfied')	✓	✓