RISE
RESEARCH ON IMPROVING SYSTEMS OF EDUCATION

RISE Working Paper 20/045
August 2020

# Identifying Effective Teachers: Lessons from Four Classroom Observation Tools

**Deon Filmer, Ezequiel Molina, and Waly Wane**

## Abstract

Four different classroom observation instruments—from the Service Delivery Indicators, the Stallings Observation System, the Classroom Assessment Scoring System, and the Teach classroom observation instrument—were implemented in about 100 schools across four regions of Tanzania. The research design is such that various combinations of tools were administered to various combinations of teachers, so these data can be used to explore the commonalities and differences in the behaviors and practices captured by each tool, the internal properties of the tools (for example, how stable they are across enumerators, or how various indicators relate to one another), and how variables collected by the various tools compare to each other. Analysis shows that inter-rater reliability can be low, especially for some of the subjective ratings; principal components analysis suggests that lower-level constructs do not map neatly to predetermined higher-level ones and suggest that the data have only a few dimensions. Measures collected during teacher observations are associated with student test scores, but patterns differ for teachers with lower versus higher subject content knowledge.

**JEL Classifications:** I20; I25; O12; O15

**Keywords:** Education; Teacher Performance; Classroom Observation

UKaid from the British people
Australian Aid
Bill & Melinda Gates foundation
Oxford Policy Management
Blavatnik School of Government
University of Oxford

**Identifying Effective Teachers: Lessons from Four Classroom Observation Tools**

Deon Filmer
World Bank

Ezequiel Molina
World Bank

Waly Wane
World Bank

Acknowledgements:

We would like to thank Diwakar Kishore for excellent research assistance in this project. Funding support from the World Bank and in particular the Knowledge for Change Program (KCP) Trust Fund, as well as from the Research on Improving Systems of Education (RISE) program are gratefully acknowledged. We also acknowledge the Tanzanian students, teachers, schools, and government officials who contributed to this research, as well as the staff at REPOA who collected the survey data. Corresponding author: dfilmer@worldbank.org.

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 Department for International Development (DFID), 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., Molina, E. and Wane, W. 2020. Identifying Effective Teachers: Lessons from Four Classroom Observation Tools. RISE Working Paper Series. 20/045. https://doi.org/10.35489/BSG-RISEWP_2020/045.

## 1. Introduction

Teacher salaries are typically the single largest investment that countries make in basic education, and evidence shows that teachers explain a significant share of variation across students' achievement (Araujo et al. 2016, Bold et al. 2019; Dobbie and Fryer 2013). At the same time, this achievement is poorly correlated with teachers' observable characteristics including age, gender, education, experience and even hours in the school (Aaronson et al. 2007; Kane and Staiger 2008; Rockoff et al. 2008). The implication is that factors such as teachers' content knowledge, pedagogical knowledge, as well as classroom behaviors and practices are potentially important determinants of students' academic performance. This paper compares various approaches to measuring these factors and explores the extent to which they are associated with student test scores—both individually and collectively.

A growing literature has documented significant shortfalls in service delivery in education across countries (World Bank 2018). Surveys of teacher absenteeism in low- and middle-income countries routinely show absence rates of between 15 and 20 percent on any given day (Chaudhury et al 2006, World Bank 2003, Das et al 2007). The Service Delivery Indicators (SDI) program (www.sdindicators.org) has documented teacher absence rates of over 20 percent in Togo and Uganda, and over 40 percent in Mozambique. In addition, even when teachers are in school, they are frequently absent from the classroom: for example, the 2014 round of SDI in Tanzania revealed that while only 15 percent of teachers were absent from the school, 47 percent of teachers (including those absent from school) were not in class at the time of an unannounced visits (Martin and Wane 2016). Even when present, teachers are devoting significant amounts of time to non-teaching and learning activities, such as classroom management (Abadzi 2009; Bruns and Luque 2015). The cumulative result is that students are receiving only a fraction of the scheduled teaching time as actual learning time: across 7 Sub-Saharan African countries, whereas scheduled teaching time was on average 5 hours and 27 minutes per school day—actual teaching time was on average 2 hours and 46 minutes (Bold et al. 2017).

In order to shed light on the measurement of these issues and their implications, we administered four different classroom observation instruments—from the Service Delivery Indicators, the Stallings Observation System, the Classroom Assessment Scoring System, and the recently developed *Teach* classroom observation instrument—in about 100 schools across four regions of Tanzania. The research design is such that various combinations of tools were administered to various combinations of teachers so these data can be used to explore the commonalities and differences in the behaviors and practices captured by each tool, the internal properties of the tools (for example how stable they are across enumerators, or how various indicators relate to one-another), and how variables collected by the various tools compare to each other. In addition to insights on the tools themselves, we leverage measures of teacher and student

test scores to investigate how the measures captured in each tool correlate with student performance. We also pool the data across all the measures to explore what we can learn from the superset of measures. The goal of this analysis is to assess the effectiveness of the instruments in capturing various dimensions of teaching practice—and to explore the relationship between those dimensions and student test scores. Importantly, the goal is not to determine whether one instrument is "better" than another. Rather, it is to understand commonalities and differences.

The analysis yields three main sets of findings. First, all the observation measures suffer from inter-rater reliability issues. Consistency across enumerators and raters tends to be highest for variables related to the share of time spent teaching and are larger for variables at higher levels of aggregation. In this implementation, the raters for *Teach* were more consistent than those for CLASS. At the same time, however, and despite the low correlation between scores assigned by the different raters, the magnitude of the differences across raters in assigned scores was relatively low.

Second, there is a limited degree of correlation between measures from the different tools even for variables that, according to their definitions, are thematically similar. At the same time, principal components analysis of lower-level "dimensions" measured by the tools suggest that these do not map neatly into the aggregate higher-level constructs as presented in the description of the different observation systems. In particular, the results suggest that while all the instruments aim to collect a number of dimensions of quality—the resulting data collected (at least in this application) have far fewer dimensions. Principal components analysis of all the variables together suggest five main dimensions: a general quality dimension; a dimension linked to good time management; a positive classroom environment (e.g. materials displayed on the walls); a dimension linked to specific instructional practices (in particular those from SDI); and a dimension linked to the material circumstances in the classroom (e.g. availability of materials and classroom infrastructure).

Third, regression analysis of student test scores suggests that, for teachers with low subject content knowledge, improvements in that knowledge and a greater share of classroom time devoted to teaching are associated with better student test scores. In addition, the results suggest that the tools are indeed identifying teacher behaviors and practices that are associated with student test scores—with different patterns for teachers with low versus high subject content knowledge. The principal component that heavily weights the CLASS variables in particular tends to be associated with better student scores for all teachers, a good classroom atmosphere is associated with higher scores for teachers with better subject knowledge, and poor support to socioemotional skills is very negatively associated with scores for teachers with low subject knowledge

The paper is structured as follows. Section 2 summarizes selected related research. Section 3 describes the sample, the study design, and basic features of the data. Section 4 presents basic statistics that emerge from the tools in this sample of teachers. Section 5 reports data on the properties of the instruments in this implementation—both their internal properties as well as how these instruments relate to each other. Section 6 explores the extent to which these instruments are associated with student test scores. A brief concluding section follows.

## 2. Selected literature review

Research from the United States has attempted to comprehensively measure and assess effective teachers and has revealed insights into the role of behaviors and practices associated with higher student learning. The Measures of Effective Teaching (MET) project showed that student performance was substantially higher when they were taught by a teacher who had been previously rated (in part based on classroom observations) as highly effective (Kane and Staiger 2012). Associated research has shown that "observation measures of teaching effectiveness are substantively related to student achievement growth and that some observed teaching practices predict achievement more than other practices" (Kane et al 2011). Importantly, much of the research carried out under the MET effort was methodological—and in particular related to classroom observations. Kane and Staiger (2012) compared five different instruments for scoring classroom instruction that had been used in the United States—Framework for Teaching (FFT); Classroom Assessment Scoring System (CLASS), Protocol for Language Arts Teaching Observations (PLATO); Mathematical Quality of Instruction (MQI) and UTeach Teacher Observation Protocol (UTOP)—and concluded that the scores on each of the five instruments were highly correlated with one another. All were associated with increases in student test scores. However, for a given teacher, scores varied considerably from lesson to lesson, and for any given lesson, scores varied from observer to observer.[2]

There has been far less systematic research on these issues in developing countries. Notable exceptions include Araujo et al. (2016) who use video recordings coded using the CLASS instrument in kindergarten classrooms in 204 schools in Ecuador, Berlinski and Schady (2015) who report on applications of CLASS in kindergarten classes in Brazil, Chile and Ecuador, Coflan, Hasan, and Raggatz (2018) who report findings CLASS instrument in 36 primary and junior secondary schools in the Guangdong province of China, Azigwe et al. (2016) who use both a low-inference and a high-inference observation instrument in 73 primary schools in the Upper East Region of Ghana, Bruns and Luque (2015) who report the findings from implementing the Stallings Observation System (hereafter referred to simply as "Stallings") approach

[2] Bacher-Hicks et al. (2019) show that teacher observations are more predictive of teacher subsequent performance than student perceptions.

to classroom observations in over 15,000 classrooms across 7 Latin American countries, Chang et al. (2014) who use video recording of 200 8th grade mathematics lessons in 200 classrooms with ex-post expert analysis in Indonesia, Seidman et al. (2018) who develop and apply the Teacher Instructional Practices and Processes System (TIPPS) instrument in 197 secondary schools in Uganda, Wolf et al (2018) who apply TIPSS in pre-primary classrooms in Ghana, the SDI surveys mentioned above which to-date have recorded observations for over 69,000 teachers in 10 Sub-Saharan African countries (www.sdindicators.org), and finally the development and application of the *Teach* instrument in Mozambique, Pakistan, the Philippines, and Uruguay (Molina et al. 2018).[3] Fewer still are studies that explicitly compare how different observation instruments perform, with Bruns, De Gregorio and Taut (2016) who compare Stallings and CLASS for 51 teachers in the Santiago Metropolitan region and two adjacent regions in Chile being a rare exception.

## 3. Sample, study design, and data

### *3.1 Setting, sample, and study design*

The study setting is Tanzania where 106 schools were reached for this Extra Teacher Observation Study (ETOS). The sample was purposively chosen from the nationally representative sample of 400 schools that had been selected for the 2014 and 2016 rounds of the Tanzania SDI survey (Martin and Wane 2016).[4] The ETOS schools are located in 11 regions across the country. Data were collected across two rounds: the first round was carried out between August and November 2016 (with almost 65% of the sample in August itself), the second round was carried out in November 2016. Eighty-eight of the schools are located in rural areas, 9 are in Dar es Salaam, and 9 are in other (semi-)urban areas (in this analysis we group these last two and designate them as "urban").[5]

The ETOS schools have on average 14 teachers and average class size of 58 enrolled students (Table 1A). In this sample (and during the first round of the survey), 14 percent of teachers were absent from the school on the day of an unannounced visit, and 39 percent of the teachers were not in the classroom (either

[3] Aslam and Kingdon (2011) use self-reported, rather than observed, practices to study the same issues.
[4] The schools were selected in a way that would minimize the additional costs (of training and enumerator visits in Round 1 and Round 2 of data collection) that would arise from adding the activities from this task to those of the 2016 round of the SDI. The resulting sample is not nationally representative and is clustered in four parts of the country: Dar es Salaam (8 schools): Western (Rukwa, Tabora, Shinyanga, Simiyu; 35 schools), Southern (Iringa, Ruvuma, Lindi, Mtwara; 36 schools), and Northern (Manyara, Kilimanjaro; 25 schools). The goal was to be roughly proportional to the number of schools in each region.
[5] The schools selected for ETOS reflect a broad geographic range, and a broad set of socio-economic conditions relevant for Tanzania. Statistical representativeness was not sought since the goal was to investigate the properties of the various instruments (internal properties, in relation to each other, and in relation to student test scores). As discussed in the text, the characteristics of schools in the sample are nevertheless similar, on average, to those in the national sample—especially within rural and urban areas.

because they were not at school or were at school but not in the classroom teaching during lesson time). There is a substantial amount of heterogeneity across rural-urban locations. Rural schools tend to be smaller with, for example, on average 11 teachers as compared to 30 in urban schools, and average class size of 54 versus 77. But even within rural or urban areas there is a large amount of variation: Schools in rural areas range from having 4 to 39 teachers, and class sizes range from 20 to 155 students—corresponding ranges in urban areas are 4 to 50 teachers, and 32 to 193 students.

The ETOS sample of schools was not designed to be representative of those across Tanzania. In particular, while only 17 percent of the ETOS schools are urban, the percentage is 48 in the national sample. Nevertheless, the mean characteristics of the ETOS schools are very similar to those from the national sample, especially after conditioning on urban/rural location. Table 1B reports the same set of characteristics as those in Table 1A but over the full nationally representative sample. Within rural schools, the ETOS schools tend to be slightly smaller and have slightly fewer students overall (they are about 10 percent lower in terms of number of teachers, students, average class size). They also have similar overall teacher absence rates. Within urban areas the characteristics are even more similar, with teacher absence from school being slightly higher in the ETOS sample (18 percent versus 14 percent in the national sample).

### *3.2 Context for classroom observations*

In order to describe the general context for this analysis, Table 2 reports summary statistics related to classroom environment and infrastructure in the sample of ETOS schools.[6] While absolute basics such as a readable blackboard and chalk seem to be common in these classrooms, more elaborate features such as functioning electricity is rare, especially in rural areas. Only around a third of classrooms have student work or other charts displayed on the walls. Availability of books for reading is extremely limited as evidenced by the almost complete lack of "corner libraries."[7]

Further to these indicators that describe physical features of the learning environment, Table 3 reports summary statistics on selected teacher practices and behaviors observed. Teachers tend to use textbooks and the blackboard, and generally engage positively with students (e.g. by standing as opposed to sitting, calling on students, or by visiting individual students). But only in about half of the classrooms were students invited to go to the blackboard, or was the teacher observed smiling, laughing, or joking with students. In about one in 5 classrooms, teachers were observed hitting a student. The vast majority of

[6] These, along with the summary statistics reported in Table 3, are derived from data collected with the SDI instrument administered during the Grade 4 classroom observations. These data are used below to create aggregates that we compare to the other instruments.

[7] Appendix 1 Table 5 reports the same means, but for the national sample. The only variable for which there is a substantive difference is the availability of electricity in the classroom: in the national sample only 4 percent of classrooms had electricity while in the ETOS sample 13 percent did.

teachers were observed engaging at least once during the lesson in what could be described as good pedagogical practices, such as asking questions that require students to apply information, use their creativity, or demonstrate understanding, but for some indicators the share is very low (for example only 21 percent of teachers summarized it at the end of the class). About two-thirds of teachers provided encouraging or "correcting" feedback (which might be considered good practice) to students. At the same time, classroom observations find few teachers providing "scolding" feedback (which might be considered bad practice) to the students. However, very few teachers assigned or reviewed homework, and few used local language for instructions. While Table 1 suggests that conditions in rural and urban schools were quite different, Table 3 suggests that teachers behaved similarly towards students in the two settings.[8]

### *3.3 Classroom observation instruments*

Four classroom observation instruments were used in ETOS; two were administered in classrooms themselves and two were used afterwards by coding videos of the lessons. During the first and second rounds of data collection, teachers were observed in classrooms using the SDI and Stallings instruments (Stallings 1976, Stallings, Knight, and Markham 2014, and World Bank 2015).[9] In addition to the in-classroom enumerators, cameras were set up in selected classrooms to video the lessons. The videos ran for the entire length of the lesson and were subsequently divided into two or three clips of around 20 minutes each. These clips were subsequently coded using both the Classroom Assessment Scoring System (CLASS; Hamre et al. 2007, Pianta et al. 2012)[10] as well as the "Teach" scoring system (Molina et al. 2018; Molina et al. 2019).[11]

Details of the instruments are described in detail below, but the key features of each is summarized in the Box 1. Three of the instruments include an explicit measurement of time-on-task (CLASS is the exception). The observation systems have different requirements in terms of enumerator or rater knowledge. The SDI and Stallings approaches are "low-inference" in the sense that they do not require enumerators to interpret much of what they are observing—rather, for these instruments, enumerators are mostly in the position of recording what they see in the form of a checklist of pre-populated categories. The CLASS and *Teach* approaches are "high-inference" since they require raters to map specific observed

[8] Appendix 1 Table 6 reports the same means, but for the national sample. The means are generally similar and within about 10 to 15 percent of each other. The largest differences are for teacher summarizes the lesson at the end of class, assigns homework, reviews homework, teacher uses local language for instruction which are roughly 50 percent lower in the ETOS sample.

[9] The SDI instrument is is available at https://microdata.worldbank.org/index.php/catalog/2748/download/39237. The Stallings instrument and manual are available at https://openknowledge.worldbank.org/handle/10986/22401.

[10] Instrument and manual are in Pianta et al. (2012). Coflan, Hasan, and Raggatz (2018) also provide information on the description and indicators related to each dimension.

[11] Instrument and manual are available at http://documents.worldbank.org/curated/en/949541542659103528/Teach-Observer-Manual.

behaviors and practices to a set of scores on various scales. Enumerator and raters for SDI, Stallings and *Teach* each received 4 days of training which included a day at a school implementing the tool. For CLASS, raters received two days of training after which they could practice independently online prior to taking a certification test.

| **Box 1: Summary of classroom observation instruments used** | | | |
|---|---|---|---|
| Instrument | Administration | Time-on-task | (Other) Areas of focus |
| SDI | In-person | Yes | Checklist of observed teacher behaviors, availability and use of materials, and classroom infrastructure. |
| Stallings | In-person | Yes | Checklist of observed use of materials. |
| CLASS | Video | No | Rater scoring across various dimensions grouped into 3 domains (Emotional support; Classroom organization; Instructional support) plus rating of Student Engagement. |
| Teach | Video | Yes | Rater scoring across various dimensions grouped into 3 areas (Classroom culture; Instruction; (promotion of) Socioemotional skills. |

***Service Delivery Indicators instrument***

The SDI instrument consists of two main parts. In the first part, an enumerator observes a full lesson and records minute-by-minute what the teacher is doing against a set of predefined activities. Activities include descriptors such as "Teacher interacts with all children as a group" or "Teacher supervises pupil(s) writing on the board" or "Teacher in class - not teaching" (see Appendix 1 Table 1 for the full list of recorded activities). Every five minutes, the enumerator also carries out a spot-check and records the number of students who are "off-task."[12] We use these data to construct two variables. First, the Share of Time Teaching, which is the share of minutes in which the teacher is recorded as engaged in teaching activities. Second, the Share of Time Teaching and Learning which is the share of time in which the teacher is engaged in teaching activities and in which no more than 6 students are "off-task." This is operationalized by multiplying the Share of Time Teaching by the share of spot-check observations in which no more than 6 students are off-task.

The second part of the SDI instrument is completed by the enumerator after the lesson is complete and consists of a series of questions on teacher behaviors and practices that were observed during the lesson, along with questions about the availability and use of materials as well as classroom infrastructure. We group these into 5 aggregates: Good Teacher Demeanor, Good Pedagogical Practices, Classroom

[12] Off-task in SDI is defined as: Chatting or interacting with other students about issues not related to the lesson; fighting, playing or having physical interaction unrelated to the lesson with other students; being disciplined; sleeping, day dreaming, or not paying attention; distracted by an activity or event inside or outside the classroom.

Environment, Availability of Materials and Classroom Infrastructure (to simplify, we refer to this as "Materials and Infrastructure"), and Use of Materials. Indexes for each of these categories are generated as the average from the (recoded) responses to the set of questions under each group (see Box 2 for a summary, and Appendix 1 Table 1 for the source variables and how they map to the aggregated indexes).

| **Box 2: Summary of variables derived from SDI classroom observation instrument** | |
|---|---|
| **Time-on-task (Level 1)** | |
| Share of Time Teaching | Share of minutes observed in which teacher is engaged in teaching activities. |
| Share of Time Teaching and Learning | Share of minutes observed in which teacher is engaged in teaching activities, adjusted by share of spot-check observations in which no more than 6 students are off-task. |
| **Teacher practices and classroom environment (Level 1)** | |
| Good Teacher Demeanor | Average of 7 0/1 variables capturing whether teachers: moved about the class; engaged with students; or projected a positive attitude. |
| Good Pedagogical Practices | Average of 13 0/1 items capturing whether teachers asked questions that stimulated thinking; provided constructive feedback to students; summarized the lesson; used homework as a tool; or uses local information from community to make learning relevant.[13] |
| Classroom Environment | Average of 2 0/1 variables capturing whether pupil work and/or other materials are displayed on the walls. |
| Materials and Infrastructure | Average of 7 0/1 variables capturing whether students have textbooks, pens, exercise books, and/or desks; whether there are reading books in the classroom; whether the classroom has a blackboard that is readable and chalk; whether the classroom has electricity; the state of hygiene in the classroom; the state of hygiene in the classroom. |
| Use of Materials | Average of 3 0/1 variables capturing whether various materials were actually used during the lesson. |
| Source: Authors. See Annex for full description of variables being aggregated. | |

***Stallings Observation System***

In the Stallings approach, an entire lesson is observed by an enumerator who records 10 "classroom snapshots" that are evenly spaced over the lesson time. At the time of each snapshot, the enumerator visually scans the room clockwise and records what the teacher is doing, what materials are being used, and how many students are engaged in that task, as well as what the students who are not engaged in that task are doing. Teacher activities include items such as "reading," "discussion," "monitoring seatwork," and "disciplining students," while other activities the other students could be doing include "social interaction" and "uninvolved" (the full set of variables recorded are in Appendix 1 Table 2). We use these data to construct two variables. First, the Share of Time Teaching, which is the share of the 10 snapshots in which the teacher is engaged in teaching activities. Second, the share of the 10 snapshots in which the teacher is

[13] Three of the variables in this group are not recorded as binary but rather as ordinal (e.g. 0=never, 1=once; 2=several times). We rescale these variables to lie between 0 and 1 prior to including them in the index.

teaching and during which there is not a "large group" or "all" students—which per the Stallings manual is defined as 6 or more students—that are off-task, namely being "disciplined by the teacher," who are involved in "social interaction" or who are "uninvolved;" we call this Share of Time Teaching and Learning. The instrument also records the availability of materials in the classroom, infrastructure and the classroom environment, from which we create two aggregates: Availability of Materials and Classroom Infrastructure (again, which we shorten to "Materials and Infrastructure"), and Classroom Environment (see Box 3; more details are in Appendix 1 Table 2).

| **Box 3: Summary of variables derived from Stallings observation system** | |
|---|---|
| **Time-on-task (Level 1)** | |
| Share of Time Teaching | Share of 10 snapshot observations in which teacher is engaged in teaching activity. |
| Share of Time Teaching and Learning | Share of 10 snapshot observations in which teacher is engaged in teaching activity and in which there are not 6 or more students who are off-task. |
| **Teacher practices and classroom environment (Level 1)** | |
| Classroom Environment | Average of 2 0/1 variables capturing whether pupil work or other materials are displayed on the walls. |
| Availability of materials and Classroom Infrastructure | Average of 5 0/1 variables capturing whether the classroom has a blackboard and chalk; whether there are reading books in the classroom; whether the classroom has electricity; whether students have textbooks/other printed material; or whether students have a notebook/writing material. |
| Source: Authors and World Bank (2015) | |

### *Classroom Assessment Scoring System*

The CLASS system involves trained raters observing a video clip and subsequently coding the totality of what they observed along 11 dimensions (these include, for example, "positive climate," "behavior management," and "analysis and inquiry"). Scores are given on a 7-point scale ranging from low (=1) to high (=7). The CLASS manual provides detailed descriptions and examples of behaviors that fit each of these scores, and to be certified each rater had to pass an exam that requires them to be within 1 point of three expert-scored videos at least 80 percent of the time. In our case, certification also required "recalibration" after every 2 weeks or 20 segments (whichever came first). This process involved coding an additional video where raters had to be within 1 point of another expert-scored video at least 80 percent of the time.[14] The scores on these 11 dimensions are then aggregated into three domains: Emotional Support, Classroom Organization, Instructional Support (see Box 4; more details are in Appendix 1 Table 3). The score for an additional domain, Student Engagement (which cuts across the other dimensions) is derived separately based on how students behave during the observation.

[14] For this exercise, native Kiswahili speakers were trained in Washington, DC USA using the CLASS protocol. All the raters were certified as CLASS raters. Videos were coded in Washington, DC, USA.

| **Box 4: Summary of variables derived from CLASS observation instrument** | | |
|---|---|---|
| Domains (Level 1) | Dimensions (Level 2) | Description |
| Emotional Support | Positive Climate | Reflects the emotional connection between the teacher and students and among students. |
| | Teacher Sensitivity | Encompasses the teacher's awareness of and responsiveness to students' academic and emotional needs. |
| | Regard for Adolescent Perspectives | Captures the degree to which the teacher's interactions with students and classroom activities place an emphasis on students' interests, motivations, and points of view and encourage student responsibility and autonomy. |
| Classroom Organization | Negative Climate | Reflects the overall level of expressed negativity in the classroom (scale reversed). |
| | Behavior Management | Encompasses the teacher's ability to provide clear behavioral expectations and use effective methods to prevent and redirect misbehavior. |
| | Productivity | Considers how well the teacher manages instructional time and routines and provides activities for students so that they have the opportunity to be involved in learning activities. |
| Instructional Support | Instructional Learning Formats | Focuses on the way in which the teacher maximizes students' interest, engagement, and ability to learn from lessons and activities. |
| | Content Understanding | The depth of the lesson content and the approaches used to help students comprehend the framework, key ideas and procedures in an academic discipline. |
| | Analysis and Inquiry | Assesses the degree to which students are engaged in higher level thinking skills through the application of knowledge and skills to novel and/or open-ended problems. |
| | Quality of Feedback | Assesses the degree to which the teacher provides feedback that expands learning and understanding and encourages continued participation. |
| | Instructional Dialogue | Content-focused discussion among teachers and students that is cumulative, with the teacher supporting students to chain ideas together in ways that lead to deeper understanding. |
| Student Engagement | Captures the degree to which all students in the class are focused and participating in the learning activity presented or facilitated by the teacher. | |
| Source: These descriptions are from Coflan, Hasan, and Raggatz (2018) which are derived from Pianta, Hamre, and Mintz (2012). | | |

***Teach classroom observation instrument***

The videos clips were subsequently also coded using the *Teach* instrument. For Teach, raters observe the 20-minute clips and code the totality of this observation according to 9 "dimensions" (these include items such as "supportive learning environment," "checks for understanding," and "social and collaborative skills"). Each dimension is scored on a 5-point scale, ranging from low (=1) to high (=5). The *Teach* manual provides detailed descriptions and examples of behaviors that fit each of these scores. After the training each rater had to pass an exam which requires them to be within 1 point of three expert-scored videos at

least 80 percent of the time.[15] The 9 dimensions are aggregated into three areas: Classroom Culture, Instruction, and Socioemotional Skills (see Box 5; further details are in Appendix 1 Table 4). The *Teach* protocol includes an additional dimension to capture time-on-task. Every five minutes within the first 15 minutes of the video clip the rater takes a snapshot of the activities and records whether the teacher is engaged in teaching activities or has provided a learning activity for most students. If yes, then the rater also records how many students are "on-task," which is reported as low (6 or more students are off-task), medium (2-5 students are off-task), and high (all students are on-task, with allowance for 1 student to be off-task).[16] We use these data to construct the Share of Time Teaching, defined as the share of the 3 snapshot observations in which the teacher is teaching or has provided a learning activity to most students, and the Share of Time Teaching and Learning as the share of the 3 snapshot observations in which teacher is teaching with no more than 6 students off-task (i.e. when students on-task is "medium" or "high").

| **Box 5: Summary of variables derived from *Teach* observation instrument** | | |
|---|---|---|
| **Time-on-task (Level 1)** | | |
| Share of Time Teaching | | Share of 3 snapshot observations in which teacher is engaged in teaching activities |
| Share of Time Teaching and Learning | | Share of 3 snapshot observations in which teacher is teaching with no more than 6 students are off-task. |
| **Teacher practices and classroom environment** | | |
| Area (Level 1) | Dimensions (Level 2) | Description |
| Classroom Culture | Supportive Learning Environment | The teacher creates a classroom environment where students can feel emotionally safe and supported. Moreover, all students feel welcome, as the teacher treats all students respectfully. |
| | Positive Behavioral Expectations | The teacher promotes positive behavior by acknowledging students' behavior that meets or exceeds expectations. Moreover, the teacher sets clear behavioral expectations for different parts of the lesson. |
| Instruction | Lesson Facilitation | The teacher facilitates the lesson to promote comprehension by explicitly articulating the objectives, providing clear explanations of concepts, and connecting the lesson with other content knowledge or students' experiences. |
| | Checks for Understanding | The teacher checks for understanding to ensure most students comprehend the lesson content. Moreover, the teacher adjusts the pace of the lesson to provide students with additional learning opportunities. |
| | Feedback | The teacher provides specific comments or prompts to help identify misunderstandings, understand successes, and guide thought processes to promote learning. |

[15] For this exercise, Tanzanian coders were trained in Bukoba, Tanzania, using the Teach Protocol. Enumerators had previously carried out Teach observations in the field for a separate exercise. All raters were certified as Teach raters. Videos were coded in Bukoba, Tanzania.

[16] In Teach, off-task is defined as follows: students who are not participating in the learning activity provided by the teacher either because they are quiet but distracted, or because they are disrupting the class. For example, in the first category, students may be staring out the window, resting their head on the desk, looking down to the floor or at the observer, or sleeping. In the second category, they may be passing notes, whispering, talking to another student during an activity that does not require talking, moving around the class, shouting, or in any other way disrupting the class.

| | | |
|---|---|---|
| | Critical Thinking | The teacher builds students' critical thinking skills by encouraging them to actively analyze content. |
| Socioemotional Skills | Autonomy | The teacher provides students with opportunities to make choices and take on meaningful roles in the classroom. Students make use of these opportunities by volunteering to take on roles and expressing their ideas and opinions throughout the lesson. |
| | Perseverance | The teacher promotes students' efforts toward the goal of mastering new skills or concepts, instead of focusing solely on results, intelligence, or natural abilities. In addition, the teacher has a positive attitude toward challenges, framing failure and frustrations as useful parts of the learning process. The teacher also encourages students to set short- and/or long-term goals |
| | Social and Collaborative Skills | The teacher encourages students' collaboration with one another and promotes students' interpersonal skills. Students respond to the teacher's efforts by collaborating with one another in the classroom, creating an environment free from physical or emotional hostility. |
| Source: Molina et al. (2019). | | |

***Structure of classroom observations in the ETOS***

As described above, the ETOS was designed in a way that the properties of the various instruments can be studied. Since one of the goals was to investigate the "internal" properties of each instrument, we chose for some classrooms to have two enumerators with the same instrument. Another goal was to compare across instruments, so we chose for some classrooms to have two enumerators with a different instrument. In Round 1, Grade 4 classrooms were assigned to either having two SDI enumerators (27 schools), two Stallings enumerators (27 schools), or one SDI and one Stallings enumerator (46 schools).[17] In addition, some classrooms were assigned to having a video camera (65 schools). In each of these schools, the two enumerators were subsequently supposed to observe an additional classroom by themselves—targeting a Grade 3 and a Grade 5 classroom. The result was a set of Grade 3 classrooms observed using SDI only (46 schools), Stallings only (45 schools), and SDI and Stallings (2 schools); and a set of Grade 5 classrooms observed using SDI only (41 schools), Stallings only (52 schools), and SDI and Stallings (2 schools).[18] Grade 3 and 5 observations did not include video cameras.

In Round 2, the process was simpler. All Grade 4 classrooms were supposed to be observed using SDI and Stallings, along with a video camera. The resulting sample includes classrooms that did indeed follow that model (84 schools). However, due to technical and coordination issues in the field, videos were not done in some classrooms (9 schools). Last, some schools were not reached and there is no Round 2 SDI,

[17] As implemented, in the data we also find one school with two SDI and one Stallings observations, and two schools with one SDI and two Stallings observations.

[18] These SDI and Stallings Grade 3 and Grade 5 classrooms were not by design, but occurred in the context of field-level decisions. It is not clear why these were done, but the most likely reason is that there was no Grade 3 or 5 classroom that could be observed by the second enumerator, so they simply joined to carry out one classroom observation.

Stallings, or video for these (11 schools). At the same time, two schools that had not been reached in Round 1 could now be reached and are included in the sample.

### *3.4 Student tests*

In addition to the classroom observations, students of the Grade 4 teachers who were observed also sat for a test in each round. The test had three sections, covering mathematics, Kiswahili, and English. The test was curriculum-based—pegged to the Grade 3 and early Grade 4 curriculum and developed by Tanzanian education academics in collaboration with the research team. The mathematics test consisted of tasks ranging from identifying and sequencing numbers, addition of one to three-digit numbers, one- and two-digit subtraction, and single digit multiplication and divisions. The Language tests consisted of several different tasks ranging from a simple task testing knowledge of the alphabet, to word recognition, to a more challenging reading comprehension test. The test in Round 1 included 15 or 16 questions for each subject.[19] Basic classical item analysis and IRT analysis suggests that the test generally worked well, although for this analysis we removed two items (2 from mathematics and 1 from Kiswahili) because of poor item functioning.[20] The Cronbach alphas for the three subjects was 0.73, 0.83, and 0.63 for mathematics, Kiswahili, and English respectively.[21] Despite having been validated against the curriculum, students generally did poorly on the test. In Round 1, students answered on average 38 percent of questions correctly in mathematics (i.e. about 5 questions), 52 percent correctly in Kiswahili (just over 7 questions), and 26 percent correctly in English (just over 4 questions).[22]

While the goal was originally to administer the tests in a way that would allow estimation of value-added models (i.e. by comparing growth in performance from Round 1 to Round 2), this is likely not advisable. The main reason is that the time between the two rounds was compressed due to delays in the implementation of Round 1. For some students, the gap between rounds was as little as just over 1 month—meaning that value added estimates would be largely meaningless. We nevertheless exploit the two rounds of data by averaging student scores across both rounds—thereby increasing the signal-to-noise ratio.

## 4. Basic descriptive statistics

[19] Round 2 included 3 additional questions for mathematics, which we ignore in this analysis.
[20] Section 2 of Appendix 2 (available online via a link on https://sites.google.com/site/decrgdeonfilmer/working-papers) details the psychometric analysis conducted on the student test score data. We make the determination of "poor item functioning" based on analyzing the results of correlation between item responses, poor properties of item characteristic curves and item information function curves.
[21] Properties are similar in Round 2: the Cronbach alphas for the three subjects was 0.85, 0.71, and 0.63 for mathematics, Kiswahili, and English respectively.
[22] Further details on the test questions and the item analysis are available in Appendix 2 (available online via a link on https://sites.google.com/site/decrgdeonfilmer/working-papers).

Before turning to an in-depth analysis of how the instruments perform and how the variables in the various instruments are related with one another, this section reports what each instrument yields in terms of descriptive statistics about what was observed in the classrooms. The four instruments provide an overlapping, but not completely overlapping, set of indicators related to what is observed in classrooms. For example, SDI, Stallings, and *Teach* provide insights into the share of time in each lesson that is spent on teaching and learning—the closest equivalent in CLASS is the Level 1 Classroom Organization variable, which includes a Level 2 variable related to "Productivity" (which, in-turn, includes the variable "maximizing learning time").

Table 4 reports summary statistics on the Level 1 variables captured by each instrument. The SDI tool suggests that teachers spend 85 percent of their time in the classroom teaching, and 75 percent of the time teaching with most of the students on-task. Teachers generally have Good Demeanor (average score of 0.67), and students have materials which are generally used (average scores of about 0.7). The mean score on Good Pedagogical Practices is 0.56, suggesting that these are not generally put into effect.[23] Classrooms tend to not have student material or other materials on the wall—leading to a generally poor scores on Classroom Environment.

The Stallings tool yields a similar set of findings—although the Share of Time Teaching and Share of Time Teaching and Learning are estimated to be lower (73 and 61 percent respectively).[24] The summary statistic for Materials and Infrastructure is lower than that for SDI (0.43 versus 0.68)—likely because SDI includes more indicators on which these classrooms score high (e.g. availability of chalk/pens/notebooks). The Classroom Environment tends to be described as poor (score of 0.28).

As discussed above, CLASS and *Teach* go into greater depth on processes within the classroom than the SDI and Stallings instruments. The CLASS instrument suggests a relatively high score in the Classroom Organization domain (5.75 on the 7-point scale), a middling score for Student Engagement (4.02), and weaker scores in the domains of Emotional Support or Instructional Support (less than 3 on both). This general pattern is not uncommon in other contexts in which CLASS has been used. In their application in primary and lower secondary schools in China, Coflan, Hasan, and Raggatz (2018) find scores of over 6 on Classroom Organization, around 4 on Emotional Support, and in the mid- to high 3s on Instructional Support. In Brazil, Chile, and Ecuador, Berlinski and Schady (2015) report that Kindergarten teacher scores

[23] As suggested by Table 3, this is driven by low prevalence of summarizing lessons at the end of a class, not using homework as a tool, and not using local language for instructions.
[24] We discuss these differences in more depth in the next section.

were at or above 4 in Emotional Support and a little higher for Classroom Organization—with Instructional Support lagging well behind (around 2).

The Teach-derived measures of Share of Time Teaching (84 percent) and Share of Time Teaching and Learning (81 percent) are close to that from SDI for the former, and higher than those for SDI and Stallings for the latter. This suggests that *Teach* "penalizes" teachers less than SDI and Stallings for students being off-task. The area-level scores suggest that these teachers score relatively highly on Classroom Culture (3.65 on the 5-point scale), average on Instruction (2.43), and relatively poorly on Socioemotional Skills (2.07). Again, this overall pattern is not unusual. For example, in primary schools in Pakistan, Molina et al (2018) find similar absolute scores (3.9, 2.3, 2.0 respectively), and a similar pattern in the Mindanao province of the Philippines (Molina et al. 2018b). The results are very close to those from the administration of *Teach* to Grade 3 teachers in a nationally representative sample of Tanzanian schools in 2019 as a part of the RISE research program (3.3, 2.6, and 2.0 respectively).

The various instruments differ in their identification of variation in the sample. Figure 1 shows the coefficients of variation (CV; the standard deviation divided by the mean) for the Level 1 variables for each of the instruments. Because it normalizes the variation of a variable in the data by its mean, the CV allows one to compare variability in a unit-free way.[25] For most of the variables in SDI, Stallings, and CLASS, CVs are similar and lie between 0.2 and 0.35. Comparing across tools, three findings stand out. First, *Teach* tends to identify less variation in the data on Share of Time Teaching or Share of Time Teaching and Learning than either SDI or Stallings. Second, the CVs for the SDI and Stallings measure of Classroom Environment are very high. Third, the CVs for the CLASS measure of Classroom Organization and the *Teach* measure of Classroom Culture stand out as being particularly low.

In Round 1, in addition to teachers from Grade 4, teachers from Grades 3 and 5 were observed using the SDI and Stallings instruments. Figure 2 illustrates how the variation across all schools (as measured by the standard deviation of the grade 4 value of the variable) compares to the variation across grades within schools (as measured by the mean, across schools, of the within-school standard deviation of each variable across grades 3, 4, and, 5). These show that while the variation across schools is always larger than that within schools, the latter is nevertheless substantial. Across these variables, the average within-school standard deviation is only 40 percent less than the across-school standard deviation (with a range from 38

[25] Note, however, that the CV is only thought to be a good measure for variables that are on a ratio scale (i.e. that have a meaningful 0). Most of the variables here are not on such a scale—we nevertheless use the CV to illustrate how the empirically measured variability across indicators measured by their standard deviation, scaled by their means, varies within and across the observation systems.

to 74 percent). This means that a lot of the difference in teacher practices and behaviors is within schools (consistent with what Bruns and Luque 2015 document for Latin America).

## 5. Properties of the SDI, Stallings, CLASS, and *Teach* instruments

We turn now to some of the properties of the various instruments. We first investigate how stable the instruments were across enumerators to assess whether some of the instruments produce more consistent results than others. Next, we analyze how the various elements of each instrument relate to each other to assess the extent to which the instruments are able to isolate independent dimensions of quality. In a last step we investigate how the elements of each instrument relate to the elements of the other instruments to assess the extent to which the instruments are identifying similar dimensions of quality.

### *5.1 How stable are the scores across different enumerators/raters?*

All four approaches have clear guidelines for how to record information, and CLASS and *Teach* have detailed instructions on what scores and scales mean, and rater's certification requires that they rate consistently with expert-rated videos. In these systems, raters nevertheless have to process what they observe and use their judgement in assigning a rating. We would perhaps therefore expect that scores assigned by different enumerators to be more similar using the low-inference approaches than the high-inference ones. In this analysis, it is important to keep in mind that the findings are not necessarily attributes of the assessment systems themselves, but rather of this implementation of them. While every effort was made to train enumerators and raters according to high standards (and all of them met the requirements associated with the observation system they were tasked with using) the findings are ultimately a reflection of both the observation system and its implementation in this context.

Table 5 reports the intraclass correlation coefficient (ICC) for Level 1 (left column) and Level 2 (right column) variables.[26] The columns also report the mean of the absolute value of the differences in the scores across the raters for the variables. The ICC values can be interpreted similarly to a standard correlation coefficient—with values closer to 1 suggesting closer agreement between the two enumerators or raters.[27] The values for SDI and Stallings are derived from the classrooms in which two enumerators observed the same teacher during the same lesson and recorded their observations in parallel (this occurred in 27 for SDI

[26] These are calculated using a one-way random effects model (implemented using the "icc" command in the Stata software package).

[27] In this and subsequent sections we describe individual correlations as being, for example, "low" and "high." There is, of course, no absolute standard for describing a correlation as such. We have used a rule of thumb that above 0.3 is notable, above 0.5 is generally high, and above 0.8 is very high.

and 28 for Stallings). The values for CLASS and *Teach* are derived from individual video clips that were separately scored by two raters (there were 44 such video clips for CLASS and 55 for Teach).

The inter-rater reliability for the SDI variables is generally high for most variables. The ICC for Share of Time Teaching is very high, 0.95, and above 0.80 for Good Teacher Demeanor and Use of Materials. This reliability decreases for Share of Time Teaching and Learning to 0.75, Good Pedagogical Practice to 0.73 and for Classroom Environment to 0.64. However, this level of concordance is not found across Level 2 variables. For example, and consistent with the gap in consistency between overall and adjusted time, enumerators are less consistent in how they code time spent with various sizes of groups of students, or in activities such as testing or maintaining discipline. At the same time, however, the inter-rater reliability for some Level 2 variables is nevertheless high: "Teacher not in class, no learning activity ongoing," "Teacher supervises pupils writing on the board," and "Teacher not in class, learning activity ongoing" have ICCs of 0.96, 0.95, and 0.93 respectively. This suggests that enumerators were able to correctly and consistently identify some specific activities, but not all. Perhaps surprisingly because it would seem to involve easily observed and verified attributes, the variable Materials and Infrastructure is low at 0.46.

This general pattern for variation of ICC for variables is similar in the Stallings results, although slightly lower for the variables related to teaching time. Inter-rater reliability for Share of Time Teaching is 0.77 and that for Share of Time Teaching and Learning is 0.63. The ICC for Classroom Environment and for Materials and Infrastructure in Stallings are 0.63 and 0.52 respectively, which are very similar to SDI. As with SDI, there is a general decrease in concordance when going to Level 2 variables.

Inter-rater reliability was quite low for the Level 1 variables in this application of CLASS—with the highest ICC of 0.20 for Classroom Organization. It is unclear why these correlations are so low, with one possibility being that it was hard for raters trained on videos of US teachers to transpose that to videos of Tanzanian teachers.[28]

ICCs are quite a bit higher for the *Teach* scores. The highest is for the Share of Time Teaching at 0.90, followed by Socioemotional Skills at 0.84, Share of Time Teaching and Learning at 0.74. For the remaining variables, the inter-rater reliability is over 0.50. The ICCs for Level 2 variables tend to be lower than those for Level 1 variables (similar to the case of SDI and Stallings) suggesting that aggregating across variables tends to reduce differences across raters.

[28] When low levels of inter-rater reliability were identified early in the coding process, several of the Tanzanian videos were master-coded using CLASS and raters were retrained using these. This does not appear to have solved the problem.

These ICCs likely overstate the extent of the disagreement between the raters: whereas the correlations are low, the actual point difference between rater scores is not large. For CLASS, for example, the mean absolute value of the difference in scores across raters is relatively small—the highest being a gap of 1.3 points (for Emotional Support) on a seven-point scale suggesting that these low correlations might be driven by scores that are similar but with a substantial amount of small random variation. This is confirmed by the fact that the share of observations for which the two enumerators assigned scores that are within 1 or within 2 points is high (Table 6).[29] For CLASS, the vast majority of raters assigned scores that were within two points of each other. The lowest degree of concordance per this metric was Emotional Support at 81.8 percent; the highest was Classroom Organization at 100 percent. The extent of concordance generally falls for the Level 2 variables, but remains above 84 percent for 10 of the 12 variables. For Teach, which is scored on a five-point scale, the concordance is even higher, especially when restricting the comparison to being within 2 points of each other, when it is always 100 percent.

As expected, the observation instruments that require less inference on the part of enumerators/raters seem to exhibit a higher degree of inter-rater reliability. At the same time, the findings suggest that the absolute value of the difference in scores across raters tends to be small, even when the correlation across raters is low.

### *5.2 What are the "internal" properties of each of the instruments?*

We use three approaches to investigate how the various components of each observation system relate to one another. First, we analyze the correlation structure across Level 1 variables; second we carry out a complementary principal components analysis (PCA) of the Level 1 variables to better understand how these might relate to overall measures of quality; third, we carry out PCA on the Level 2 variables to investigate whether they tend to map well to groupings set out by the observation systems themselves (i.e. the Level 1 variables).

#### *Correlations among Level 1 variables*

Table 7 reports the cross-variable correlations within each observation system. These correlations are calculated in each case for the full sample in which each instrument was implemented (which is why the sample sizes vary across instruments). Each observation in this sample corresponds to an observed full lesson. If there was more than one enumerator/rater for a lesson or a video clip then these scores were averaged in this analysis (and the analysis that follows). In the case of CLASS and Teach, scores have been

[29] Recall that in order to be certified as a rater for CLASS or *Teach* raters have to score within 1 point of expert-rated video clips 80 percent of the time.

aggregated to the level of the lesson—so if there were two video clips corresponding to different segments then the scores on these were averaged for this analysis (and the analysis that follows).

In the SDI instrument, high levels of time spent teaching and learning tend to be highly correlated with Good Teacher Demeanor, the use of Good Pedagogical Practices, and the Use of Materials in the lesson. This suggests that these particular "good" behaviors seem to move together (or, at least, are recorded as doing so). As in the Stallings instrument, there is little or no correlation between these behaviors and the measures of the availability of materials or classroom conditions.

The various Level 1 variables captured in the CLASS instrument are highly correlated with one another, with the highest correlation being between Emotional Support and Instructional Support (0.81). Emotional support also correlates strongly with Overall Class Score (0.94). There is also be a high level of correlation between these three Level 1 variables and Student Engagement.

In the *Teach* instrument, there is a relatively low (albeit positive) correlation between the various Level 1 variables. The correlation between the Share of Time Teaching, or Share of Time Teaching and Learning, and the other Level 1 *Teach* variables are very low (perhaps surprising since was not the case for SDI). Nevertheless, given their general thematic similarity, it is perhaps comforting that the highest correlation among these is Share of Time Teaching with Instructional Support at 0.12. The highest correlation among the other variables is that between Socioemotional Skills and Classroom Culture at 0.34.

***Principal components analysis of Level 1 variables***

The correlations discussed above suggest that the tools behaved differently in terms of the extent to which the various dimensions identified are correlated with each other. In particular, the Level 1 variables in CLASS were highly correlated, while those for *Teach* were not. We extend this analysis by carrying out PCA analysis on these variables to assess how many dimensions are identified. The left-hand columns of Table 8 report the summary statistics—namely the Eigenvalues, the difference in the Eigenvalues, and the proportion of the variance-covariance—for the first three principal components for each set of Level 1 variables. The right-hand columns of Table 7 report the component loadings for the first three components for each set.[30] If one were to use the rule of thumb approach of retaining only components whose Eigenvalue is greater than 1 then none of these observation systems could be said to identify more than 2 "dimensions" of quality, and CLASS identifying only 1.[31]

[30] Note that the output includes higher order components, but for compactness we only report the first three here.
[31] In this analysis we include only the Share of Time Teaching and Learning (and not the Share of Time Teaching). This is because, by construction, the two are highly conceptually and statistically correlated and including both

There are 2 Eigenvalues greater than 1 for SDI—and a visual inspection of the component loadings suggests that these are "teacher behaviors" (which cover time use, Teacher Demeanor, Pedagogical Practices, and Use of Materials), and "physical environment." Stallings also identifies two components, the first of which is similarly mostly related to time use and Materials and Infrastructure, and the second of which relates to Classroom Environment. There is only 1 Eigenvalue in the analysis of CLASS Level 1 variables that exceeds 1—with a very large drop-off between the first and second, and similar component loadings across the variables for the first principal component—which suggests that, in essence, CLASS is identifying one dimension of quality. While the Eigenvalue for the second principal component for CLASS is less than one, the component loading structure is striking in that this component heavily loads positively on Classroom Organization and negatively on the other variables. Last, *Teach* identifies two components, with the component loading suggesting that the first is classroom practices as a group (Classroom Culture, Instructional Support and Socioemotional Skills), and that the second is largely time use (with a smaller positive component loading on Instruction).

In sum, these results suggest that while all the instruments aim to collect a number of "dimensions" of quality—the resulting data collected (at least in this application) have far fewer dimensions.

***Principal components analysis of Level 2 variables***

We next carry out a similar analysis but on the Level 2 variables for each observation system. The main interest here is to assess whether the PCA component loadings "recover" the various conceptual distinctions the systems have for their Level 1 variables. For SDI and Stallings this exercise is not very informative since the main difference in the Level 1 and Level 2 variables is a finer disaggregation of the time variable. Since the total time is fixed (at 100 percent), these variables are constrained in the way they can move together. We nevertheless report these results for completeness, but refrain from trying to interpret them. The results can be found in Table 9.

For CLASS the Eigenvalues and component loadings suggest at most 2 dimensions, the first being an overall positive quality dimension, and the second loading positively on those Level 2 variables that map to Emotional Support and Classroom Organization and loading more negatively on those Level 2 variables that map to Instructional Support. For Teach, the PCA analysis suggests up to 4 dimensions—although these do not map neatly to the time use plus the three pre-specified dimensions of Teach. The component loadings for the first component suggests that this is an overall positive quality dimension that loads relatively equally across the Level 2 variables. The second component loads positively on Feedback,

makes results hard to interpret. If we include both in these models, they both have very similar factor loadings in the first component, and other results are qualitatively unaffected.

Critical Thinking, and Checks for Understanding. All these Level 2 variables map to the Instruction area, indicating a dimension linked to that construct. The third component loads positively on Share of Time Teaching and Learning, along with variables such as Positive Behavioral Expectations (which maps to Classroom Culture) and Lesson Facilitation (which maps to Instruction). The fourth component loads positively on Share of Time Teaching and Learning, Positive Behavioral Expectations (which maps to Classroom Culture), and Social and Collaborative Skills (which maps to Socioemotional Skills).

For each observation system, the first Eigenvalue is substantially larger than the second, suggesting that there is typically one predominant dimension of quality that emerges from the Level 2 variables. To the extent that more than one dimension is identified, these dimensions do not map directly into the concepts identified by the Level 1 variables.

### *5.3 How do the Level 1 variables relate to one another across instruments?*

A key feature of these data is that they allow us to compare how the different instruments relate to one another. We explore these relationships using two approaches: first by correlating the various Level 1 variables with each other; second by carrying out principal components analysis on all of them together.

#### *Correlational analysis*

Table 10 reports the full set of correlations across Level 1 variables for each pair of instruments. These correlations are carried out on data that have been aggregated to the lesson level (i.e. the scores from two video clips from different parts of the same lesson have been combined) and averaged across multiple enumerators/raters when there are more of one of these for the same lesson or video clip. In each case the sample includes all observations in which the lesson has a score from the two instruments in question, with the implication that the different correlations (e.g. SDI vs. CLASS and CLASS vs. Teach) are not always over the same sample.

The most remarkable feature of the correlation coefficients reported in Table 10 is how low they are. Across the three instruments that measure Share of Time Teaching, the correlations across instruments for these variables is highest for that between SDI and Stallings at 0.80 (Panel A), 0.50 for the correlation between SDI and *Teach* (Panel C), and as low as 0.43 for that between Stallings and *Teach* (Panel E).[32] The correlations between the Share of Time Teaching and Learning are generally lower.

[32] Statistical significance of these correlation coefficients is available from the authors on request. In general, for these results, correlation coefficients above 0.15 are statistically significantly different from zero at the 5 percent level.

To illustrate how these measures differ, Figure 3 plots the density distributions of the various indicators of Share of Time Teaching (Panel A), and Share of Time Teaching and Learning (Panel B). To help isolate differences, the third panel (Panel C) plots the densities of the difference between the two, namely the time lost due to students being off-task.[33] These distributions, while consistent with the averages (Table 4) and correlations (Table 10) for these variables, nevertheless provide additional insights. SDI, which records time in a minute-by-minute fashion, has the right-most distribution for Share of Time Teaching—meaning it assigns a greater share of time to be classified as teaching. In addition, the adjustment for learning based on the share of time with no more than 6 students off-task (which is recorded during snapshots at 5-minute intervals) has the left-most distribution. Stallings, where time on teaching is recorded during a snapshot every 10 minutes, has the left-most distribution for Share of Time Teaching—meaning that it assigns the least time as being classified as teaching. The distribution of time lost due to students being off-task is the right-most of the three tools—meaning that it is most likely to record an observation as being a non-learning one. Last, the *Teach* instrument, which measures the time teaching in 3 snapshots during the first 15 minutes of a video clip, has an overall distribution of Share of Time Teaching time that is similar to that from Stallings. At the same time, however, the adjustment in *Teach* has a distribution that is somewhere between that of SDI and Stallings—with few high-values for share of time lost.

On net, Stallings and *Teach* have similar distributions in terms of Share of Time Teaching—and both are different from SDI. But patterns change after adjusting for off-task students. The adjustment has different effects for Stallings versus Teach, and the distributions are no longer similar. On the other hand, the distributions for SDI and *Teach* become similar after adjusting.

Teaching and learning time from the various sources is generally positively associated with all four dimensions measured in CLASS. While the magnitudes of the coefficients are not generally large (typically around 0.2 to almost 0.5), and they are consistently statistically significantly different from zero. The correlation coefficient is highest in the case of the "Classroom Organization" variable in CLASS (where it reaches 0.55 for the correlation with Share of Time Teaching from SDI, Panel B). Share of Time Teaching and Share of Time Teaching and Learning are modestly (albeit statistically significantly) correlated with the Instruction variable from *Teach* (Panels C and E).

The correlations between the high-inference variables in CLASS and *Teach* do not suggest a close mapping between any of these (Panel F). The variables are all positively correlated with one another (with coefficients that are statistically significantly different from zero), but small: the highest correlation

[33] Figure 3 shows the distributions using all observations for each instrument. Restricting the sample to observations for all three instruments yields very similar results (see Appendix 1 Figure 1).

coefficient is between Instructional Support from CLASS and Instruction from *Teach* (0.31). The aggregate measures derived from these two instruments are correlated with an overall correlation coefficient of 0.36.

***Principal components analysis***

Table 11 reports the results of a principal components analysis of all the Level 1 variables simultaneously. Note that this analysis can only be carried out on the 107 observations (lessons) where we have data for all four instruments. The Eigenvalues suggest that the data contain a number of dimensions; seven of them are greater than 1. At the same time, the difference between the first and the second is large, again suggesting that the main thrust of these variables is captured in the first principal component. The component loadings for this first component are generally positive and mostly range between 0.15 and 0.42 for the behaviors and practices variables, so this component could be characterized as "overall good teaching practices" (the component loadings are highest for the CLASS variables).

Proving an interpretation to the other principal components is more difficult. The second component loads most heavily on variables linked to the various Share of Time Teaching and Learning variables, as well as to the CLASS variable of Classroom Organization—suggesting that "good time use" is one dimension of the data. Component 3 loads heavily on the Classroom Environment variables from SDI and Stallings, suggesting "good classroom environment" as a third dimension (the Classroom Environment variable is built from measures of student work or other materials being displayed on the walls). Component 4 seems to be recovering the teacher practices as captured in SDI (loading positively on Good Teacher Demeanor, Good Pedagogical Practices, and Use of Materials) as well as more modestly on Instruction from Teach—suggesting a dimension related to "good instruction" that is not captured in the first component. The fifth component loads mostly on Materials and Infrastructure (from SDI and Stallings) along with Classroom Culture from Teach—suggesting it is capturing something along the lines of "material circumstances."

The main finding to emerge from this is that the data sort themselves somewhat neatly into dimensions that one might expect. First, a general quality dimension; second a dimension linked to good time management; third, a positive classroom environment (with materials displayed on the walls); fourth, a dimension linked to instructional practices (at least those not directly captured in the first component); and fifth, a dimension linked to the material circumstances in the classroom.

**6. How do measures from these instruments correlate with student test scores?**

The final step in this analysis is to investigate the degree to which the various dimensions of teacher behaviors and practices identified in these observation systems are related to student performance. As

discussed in Section 1, Grade 4 students in these schools were administered two rounds of tests (each covering the subjects of mathematics, Kiswahili, and English) and we use these to calculate an overall score averaging over subjects and the two rounds.[34] Specifically, we estimate two-parameter IRT models for each subject and each round separately. We then extract the latent ability parameter from each model, average across these, and normalize the resulting variable so it has mean 0 and standard deviation 1.[35]

We estimate a series of regression models with the student test score as the dependent variable against various combinations of the variables that emerge from the observation instruments. We include two variables related to the share of time spent in teaching and learning: the Share of Time Teaching and a separate variable equal to the share of time lost due to students being off-task.[36] In addition, we include a test score for the teachers' subject content knowledge. This is derived from a test administered as a part of the national SDI survey structured as a teacher correcting mock student tests in language and math. The test covered material at all grades of primary school.[37] The score we use is based on an item response theory analysis of these data.[38]

We estimate each model first without controlling for other variables, and second including a set of student, household, teacher, and school characteristics. Student and household characteristics include age, gender, whether the student had eaten before school on the day of the test, and a number of characteristics that reflect socioeconomic status (having a separate room to sleep in at home; having electricity at home; having running/tap water at home; living in a dwelling with concrete/cement/stone walls; living in a dwelling with a metal roof; living in a dwelling with a toilet; household ownership of various assets—bed, mosquito net, books, mobile phone, computer).[39] Teacher characteristics include variables reflecting gender, age, education, and training. School characteristics include indicators for whether the school has a way of "recognizing" teacher performance, availability of piped water at the school, accessibility to a road, the ratio of students to teachers, and the location of the school (urban/rural). We also control for the subject being taught during the observation.

[34] Results disaggregated by subject are qualitatively similar (these are available online via a link on https://sites.google.com/site/decrgdeonfilmer/working-papers).
[35] We also estimated results using a Rasch model, as well as simply the percent correct and results are qualitatively similar (these are available online via a link on https://sites.google.com/site/decrgdeonfilmer/working-papers).
[36] This is calculated as Share of Time Teaching minus Share of Time Teaching and Learning.
[37] See Bold et al. (2019) for further description of the teacher test.
[38] The specific measure we use comes from an analysis of pooled data from a number of Sub-Saharan African countries.
[39] In a number of cases, some of these variables are missing. In such cases we replace the value by the mean across the sample, and include a dummy variable in the model that is equal to 1 if the value was originally missing.

For each observation system, and then for all of them combined, we first estimate a model with all Level 1 variables and then a model with the principal components that emerge from a principal components analysis of them.[40] Last, in order to investigate heterogeneity, we separate the sample into teachers who score below and above 0 on the (normalized) teacher test (we refer to this below as the "threshold").[41] The coefficient estimates on the teacher observation and teacher score variables from all of these regressions are reported in Tables 12 to 16.[42] The tables also report p-values for F-tests of whether the coefficients on the variables in each of the sets of control variables are jointly equal to zero.

The data we use for this analysis are slightly different from those used above to investigate the properties of the various instruments. That analysis was carried out at the classroom observation level, whereas this analysis is carried out at the teacher level. Specifically, we aggregate all information from each observation tool to the level of a teacher. For example, if a teacher was observed with the SDI tool in Rounds 1 and 2, we average across the two rounds; or if a teacher has two videos that were coded using *Teach*, then we average across those. This teacher level file is then merged with the student test scores and it is on this data that we run the regression analysis (and report standard errors that cluster at the teacher level). In order to ensure comparability across the models, we estimate these regressions on the sample for which we have all four observation types.

It is important to recognize that we cannot provide a causal interpretation to these estimates. While controlling for student, family, teacher, and school characteristics might help to identify the link from teacher skills and behaviors to student test scores, there are a variety of potential selection and reverse causation issues that we cannot rule out. The findings are therefore only indicative of (conditional) associations in the data.

### *6.1 Instrument-by-instrument results*

[40] Note that this PCA analysis is slightly different from that reported in Section 5. First, that analysis was done with all individual observations; this analysis is done at the level of each teacher—with all observations for that teacher with a particular instrument averaged. Second, in this analysis we exclude teaching and learning time from the principal components analysis and treat it as a qualitatively different type of variable. In addition, we remove redundant variables and so, for example, include only one variable for share of time on teaching and learning. The results from this analysis are reported in Appendix 1 Tables 6 and 7 and are discussed further below.

[41] The mean score across teachers is -0.05 and the standard deviation is 0.63 (the median score is 0.03).

[42] We also estimate models where each variable enters one at a time since there is a degree of correlation between them and associations might get masked in the multivariate regression context. The results of these variable-by-variable models (first including only dummy variables for the subject observed, and then including for all the control variables) are reported in Appendix 1 Table 9. The results are generally consistent, although in the case of CLASS more variables emerge as being individually significantly associated with student test scores than in the multivariate models. This is consistent with the fact that they were highly correlated with each other, as discussed in Section 5.

The results do not suggest that the teacher variables for any of the observations systems alone (along with the test score) capture a large share of the overall cross-sectional variation in student test scores: the share of variation (as captured in the R-square) ranges between on the order of 2 to 6 percent. Splitting the sample into teachers who score below and above the threshold of 0 on the teacher test improves this share: in these disaggregated models it ranges from about 5 to 17 percent. Including all the variables together (Table 16) increases the share in the pooled model to between 8 and 11 percent, and in the split models to between 10 and 20 percent. The fact that including all the variables together increases the explanatory power suggests that the different tools are indeed picking up different dimensions of teaching quality—and that together they can explain up to 20 percent of the variation in student scores (albeit, only in particular samples). Adding in the control (student, teacher, school) variables boosts the overall share of variation explained by about an additional 20 percentage points.

The main finding to emerge from these regression models is the difference in patterns across teachers who scored below and above the threshold on the teacher test. For teachers who scored below 0, the teacher test score is generally positive and consistently statistically significantly associated with student test scores (Columns 5 to 8 of Tables 12 to 15). Excluding the Stallings models, the average magnitude of the estimated coefficients is 0.19 (ranging from 0.12 to 0.26 across the models).[43] Going from a teacher at the 10$^{th}$ percentile of teachers in this group to the 90$^{th}$ percentile (an increase of 0.91 in the teacher test score) is therefore associated with an incremental 0.17 standard deviations on the student test score (with a range of 0.11 to 0.23 standard deviations depending on the model). In contrast, for teachers above 0 this association is no longer statistically significant. It is of note that the variation is teacher scores is substantially larger in the group below the threshold (standard deviation of 0.59) than in the group above (standard deviation of 0.26), and that the distribution has a particularly long lower tail. These results are therefore likely driven by very low-quality teachers (in terms of subject content knowledge) being associated with poorly performing students.

The results on the Share of Time Teaching, and the share of time lost to students being off task, do not have as consistent a pattern across the models. The results do, however, suggest that teaching time—as measured in SDI and Teach—is positively associated with student learning among teachers below the threshold. For example, in the SDI results (Table 12) Share of Time Teaching is statistically significantly associated with student test scores in all the models for teachers below the threshold. Across these models the estimates imply that going from a teacher at the 10$^{th}$ percentile to one at the 90$^{th}$ percentile (in terms of

[43] Stallings seems to be the exception here, with positive but non-statistically significant coefficient estimates. It is unclear why this is the case—but is perhaps linked to the fact that the Share of Time Teaching is also not significant in these models, unlike the case of SDI and Teach.

share of time spent teaching—this difference is 0.33) in this group is associated with student test scores that are between 0.28 and 0.69 standard deviations higher (for the 25th to 75th percentile where the difference is 0.15, the range is 0.13 to 0.31 standard deviations). This pattern and the magnitudes are similar using the *Teach* version of this variable (Table 15). For teachers above the threshold, the results are less consistent. They are smaller in magnitude and only statistically significant in some specifications for SDI, and never statistically significant for Teach. The models that use the Stallings version of this variable produce very small and statistically insignificant coefficients.

The regressions do not strongly support the notion that the share of time lost is highly negatively associated with student test scores—although the associations are generally negative for SDI and Stallings, and for teachers above the threshold for Teach. For Stallings, the estimates are negative, large, and statistically significant for teachers below the threshold.

The other variables obtained through the various observation instruments do not appear to exhibit a systematic pattern in terms of their association with student test scores. Using the SDI and Stallings instruments, the results point to an association with Availability of Materials, especially for teachers above the threshold (Tables 12 and 13). For CLASS, the results point to Classroom Organization and Instruction as being associated with student scores for teachers above the threshold (Table 14). For Teach, the results point to Instruction as being positively associated with student scores for teachers above the threshold and, perhaps surprisingly, negatively so for teachers below the threshold (Table 15). Last, the *Teach* variable of Socioemotional Skills is positively and statistically significantly associated with student test scores, but only for teachers below the threshold.

The models that include the principal components derived from each set of variables are generally consistent with those that include individual variables, in the sense that the former (principal components) are generally statistically significantly associated with student scores when one or more of the latter variables are.

### *6.2 All instruments together*

In order to simultaneously explore the full set of variables captured in these observation instruments we estimate models that include all of them (Table 16).[44] The pattern of results on teacher test scores is confirmed in these models: they are large and statistically significantly positively associated with test scores for students of teachers below the threshold, and insignificant for teachers above. The Share of Time

[44] We remove redundant variables in this exercise, so for example, we only include the Share of Time Teaching, share of time lost to students being off-task, Availability of Materials, and the Use of Materials variables from the SDI instrument.

Teaching is large and positively associated with student scores for teachers below the threshold. For teachers above the threshold is generally statistically significant, although the size of the association is smaller (the coefficients fall by about 75 percent, from around 2 to around 0.5).

Variables that emerge as positively associated with student scores for teachers below the threshold are Student Engagement from CLASS and Socioemotional Skills from Teach. For teachers above the threshold the variables that are positively associated with student scores are: Availability of Materials from SDI and Instruction from Teach.

Of course, the coefficient estimates discussed above are all conditional on one another, making it hard to clearly interpret them (and, in particular, the negative estimates). Aggregating all the variables using an approach such as principal components can therefore be useful, since it identifies a set of orthogonal dimensions of the data. However, interpreting the results from the models that use principal components requires an attempt to interpret the pattern of component loadings from those models (recall that this is slightly different from the principal components analysis to in Section 5 because here we aggregate to the level of the teacher and exclude the time variables from this exercise). Visual inspection of these loadings (Appendix 1 Table 8) suggests the following interpretations.

The first principal component loads fairly evenly across the teacher behavior variables from the various instruments (component loadings are largest for the CLASS variables) and could therefore be characterized as "general good teaching practices." The second component loads most positively on the SDI teacher behavioral variables (Good Teacher Demeanor, Good Pedagogical Practices, Use of Materials), but negatively on three of the CLASS behaviors (Emotional Support, Instructional Support, Student Engagement) and could therefore be characterized as "good SDI/bad CLASS." The third principal component loads most heavily on aspects of classroom organization (Classroom Environment, Materials and Infrastructure from SDI, Classroom Organization from CLASS, Classroom Culture from Teach) and could therefore be characterized as "good classroom atmosphere." The fourth component is perhaps most marked by a large negative component loading on Socioemotional Skills, and a slightly smaller negative loading on Classroom Culture, both from Teach, along with a positive loading on Materials and Infrastructure from SDI; the component could perhaps be characterized as "poor support to socioemotional skills development."

Consistent with previously discussed results, teacher test scores are significantly associated with student scores for teachers below the threshold in these models. Higher values of "Good SDI/Bad CLASS" are negatively associated with student scores which is consistent with an interpretation that CLASS is better than SDI at capturing those teacher practices that are positively associated with student learning. "Good

classroom atmosphere" is associated with higher scores, but only for teachers above the threshold. Last, "poor support to socioemotional skills" has a large negative association with student scores for teachers below the threshold. Going from a teacher at the 90$^{th}$ percentile of the distribution of this variable in this group of teachers to one at the 10$^{th}$ percentile (a difference of 2.4) is associated with an increment of 0.52 standard deviation in student scores.

In sum, the most consistent finding to emerge from this analysis is that for teachers with low subject content knowledge, improvements in that knowledge and a greater share of classroom time devoted to teaching are associated with better student test scores. At the same time, the results suggest that the tools are indeed identifying teacher behaviors and practices that are associated with student test scores. The principal component that heavily weights the CLASS variables in particular tends to be associated with better student scores for all teachers, a good classroom atmosphere is associated with higher scores for teachers with better subject knowledge, and poor support to socioemotional skills is very negatively associated with scores for teachers with low subject knowledge.[45]

## 7. Conclusion

In this study we have implemented four different teacher observation instruments—SDI, Stallings, CLASS, and Teach—in a sample of about 100 schools in Tanzania. There are three main sets of findings. First, the results suggest that inter-rater reliability is not always very high. In the case of CLASS, inter-rater reliability was low, although the absolute magnitude of the differences across raters was small. This suggests that in any application, care needs to be taken in enumerator training to ensure consistency in observations and ratings.

Second, measures associated with time-on-task tend to be correlated across instruments. We do not find, however, that the other variables from the various instruments are highly correlated with each other, even when they would appear to be thematically similar. At the same time, principal components analysis suggests that the variables collected seem to organize themselves into dimensions of quality that generally cut across various instruments. The leading component suggests that together, the observed behaviors and practices capture an overall quality measure (that is, component loadings are fairly consistent across the variables). This suggests that, as a practical matter, the specific instrument used may not be of overriding importance if the goal is to capture an overall sense of the "quality" of teachers as measured by classroom behaviors. For high-frequency monitoring in the context of regular quality assurance or inspectorate

[45] The findings are very similar if we use a "percent correct" approach to measuring student scores, or if we use a 1PL model for the IRT derivation of latent ability. Results are available online via a link on https://sites.google.com/site/decrgdeonfilmer/working-papers.

functions of the ministry, it might be necessary to pare-down the instruments to a more limited set of indicators. These could be expanded or changed as areas of focus for improvement change.

Last, none of the instruments produces measures that are individually highly predictive of student test scores. At the same time, they do produce measures that are associated with student test scores, In the subsets of teachers with either low or high subject content knowledge, the measures explain 15 to 19 percent in the variation of student test scores. This suggests that the instruments capture aspects of teacher quality that may matter for student learning outcomes. At the same time, more research is needed to establish the causal nature of that association, as well as which specific behaviors—beyond time spent on teaching and learning—might be driving it.

Despite the richness of these data, the analysis nevertheless has limitations. An overarching one is that the findings discussed here are not necessarily attributes of the instruments themselves, but of the instruments and this implementation. With different training protocols, or in a different context, results could be different. Research that carries out similar analysis in different countries could help to shed light on the extent to which these findings are consistent across settings. An additional limitation is that the student test scores we use to explore the relationship between teacher practices and behaviors and student learning outcomes are measured cross-sectionally as we could not confidently establish a measure of teacher value-added. Such data would have led to a more causal interpretation of the associations between observed teacher behaviors and student learning outcomes. Last, for some comparisons the number of data points is limited (in particular, the inter-rater reliability analysis is based on samples that range from only 27 to 55 observations). Further research that repeats this type of analysis with larger samples would help address this issue.

Understanding which teacher behaviors and practices most closely map to better student learning outcomes, and how to measure those behaviors and practices, are important steps to designing better policies and programs for recruiting and training teachers. More experience with the various tools described here will be a key part of that process. If implemented in a way that results can be directly compared across tools, this experience will shed further light on how to overcome the measurement challenges involved.

**References**


Aaronson, D., L. Barrow, and W. Sander. 2007. "Teachers and Student Achievement in the Chicago Public High Schools." *Journal of Labor Economics* 25(1): 95-135.

Abadzi, Helen. 2009. "Instructional time loss in developing countries: Concepts, measurement, and implications." *The World Bank Research Observer* 24(2): 267-290.

Araujo, M. Caridad, Pedro Carneiro, Yyannú Cruz-Aguayo, and Norbert Schady. 2016. "Teacher quality and learning outcomes in kindergarten." *The Quarterly Journal of Economics* 131(3): 1415-1453.

Aslam, Monazza, and Geeta Kingdon. 2011. "What can teachers do to raise pupil achievement?" *Economics of Education Review* 30(3): 559-574.

Azigwe, J. B., Leonidas Kyriakides, Anastasia Panayiotou, and Bert PM Creemers. 2016. "The impact of effective teaching characteristics in promoting student achievement in Ghana" *International Journal of Educational Development* 51: 51-61.

Bacher-Hicks, Andrew, Mark J. Chin, Thomas J. Kane, and Douglas O. Staiger. 2019. "An experimental evaluation of three teacher quality measures: Value-added, classroom observations, and student surveys." *Economics of Education Review* 73 (): 101919.

Bruns, Barbara, and Javier Luque. 2015. Great Teachers: How to Raise Student Learning in Latin America and the Caribbean. Washington, DC: World Bank.

Barbara Bruns, Soledad De Gregorio and Sandy Taut. 2016. "Measures of Effective Teaching in Developing Countries." RISE Working Paper 16/009. https://riseprogramme.org/sites/www.riseprogramme.org/files/publications/RISE_WP-009_Bruns_0.pdf

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

Bold, Tessa, Deon P. Filmer, Ezequiel Molina, and Jakob Svensson. 2019. The Lost Human Capital: Teacher Knowledge and Student Achievement in Africa" World Bank Policy Research Working Paper No. 8849. The World Bank.

Bold, Tessa, Deon Filmer, Gayle Martin, Ezequiel Molina, Brian Stacy, Christophe Rockmore, Jakob Svensson, and Waly Wane. 2017. "Enrollment without learning: Teacher effort, knowledge, and skill in primary schools in Africa." *Journal of Economic Perspectives* 31(4): 185-204.

Chang, Mae Chu, Sheldon Shaeffer, Samer Al-Samarrai, Andrew B. Ragatz, Joppe de Ree, and Ritchie Stevenson. 2014. *Teacher Reform in Indonesia: The Role of Politics and Evidence in Policy Making*. Directions in Development. Washington, DC: World Bank

Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics* 18(1), 5–46.

Chaudhury, Nazmul, Jeffrey Hammer, Michael Kremer, Karthik Muralidharan, and F. Halsey Rogers. 2006. "Missing in Action: Teacher and Health Worker Absence in Developing Countries." *Journal of Economic Perspectives* 20(1): 91–116.

Coflan, Andrew Moore, Andy Ragatz, Amer Hasan, and Yilin Pan. 2018. "Understanding Effective Teaching Practices in Chinese Classrooms: Evidence from a Pilot Study of Primary and Junior Secondary Schools in Guangdong, China." World Bank Policy Research Working Paper No. 8396. Washington, DC.

Das, Jishnu, Stefan Dercon, James Habyarimana and Pramila Krishnan. 2007. "Teacher Shocks and Student Learning: Evidence from Zambia." *The Journal of Human Resources* 42(4):820-862.

Dobbie,W., and Fryer Jr, R. G. (2013). Getting beneath the veil of effective schools: Evidence from New York City. *American Economic Journal: Applied Economics* 5(4): 28–60.

Halpin, Peter F., and Michael J. Kieffer. 2015. "Describing profiles of instructional practice: A new approach to analyzing classroom observation data." *Educational Researcher* 44(5): 263-277.

Hamre, Bridget K., Robert C. Pianta, Andrew J. Mashburn, and Jason T. Downer. 2007. "Building a science of classrooms: Application of the CLASS framework in over 4,000 US early childhood and elementary classrooms." *Foundation for Childhood Development*.

Hanushek, Eric A. and Ludger Woessmann. 2008. "The Role of Cognitive Skills in Economic Development." *Journal of Economic Literature* 46(3): 607-68.

Hanushek, Eric A. and Ludger Woessmann. 2015. *The Knowledge Capital of Nations*. Cambridge: The MIT Press.

Kane, Thomas J., Eric S. Taylor, John H. Tyler, Amy L. Wooten. 2011. Identifying Effective Classroom Practices Using Student Achievement Data." *Journal of Human Resources* 46(3): 587-613.

Kane, T. J., and D. O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Validation." NBER Working Paper No. 14607.

Kane, Thomas J., and D.O. Staiger. 2012. Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Bill & Melinda Gates Foundation.

Martin, Gayle and Waly Wane. 2016. "Education Service Delivery in Tanzania." World Bank Report No. AUS5510. https://microdata.worldbank.org/index.php/catalog/2748/download/39242

Metzler, J., and Woessmann, L. (2012). "The impact of teacher subject knowledge on student achievement: Evidence from within-teacher within-student variation." *Journal of Development Economics* 99(2), 486–496.

Molina, Ezequiel; Syeda Farwa Fatima, Andrew Y C Ho, Carolina Melo Hurtado, Tracy Wilichowksi, Adelle Pushparatnam. 2018. "Measuring Teaching Practices at Scale : Results from the Development and Validation of the *Teach* Classroom Observation Tool." World Bank Policy Research Working Paper No. 8653. Washington, D.C.

Molina, Ezequiel , Syeda Farwa Fatima, Iva Trako, and Tracy Wilichowski. 2018b. "Teacher Practices in Mindanao: Results of the *Teach* Classroom Observation Study." Unpublished Manuscript. The World Bank.

Molina, Ezequiel; Carolina Melo Hurtado, Adelle Pushparatnam, and Tracy Wilichowski. 2019. *Teach: Observer Manual*. Washington, D.C.: World Bank

Montenegro, Claudio E. and Harry A. Patrinos. 2014. "Comparable Estimates of Returns to Schooling Around the World." World Bank Policy Research Working Paper No. 7020.

Pianta, Robert C., Bridget K. Hamre, and Susan Mintz. (2012). *Classroom assessment scoring system: Upper Elementary Manual*. Teachstone

Rockoff, J.E., B. A. Jacob, T.s J. Kane, and D. O. Staiger. 2008. "Can You Recognize An Effective Teacher When You Recruit One?" NBER Working Paper 14485.

Seidman, Edward, Sharon Kim, Mahjabeen Raza, Miyabi Ishihara, and Peter F. Halpin. 2018. "Assessment of pedagogical practices and processes in low and middle income countries: Findings from secondary school classrooms in Uganda." *Teaching and Teacher Education* 71: 283-296.

Stallings, Jane A. 1976. "How instructional processes relate to child outcomes in a national study of follow through." *Journal of Teacher Education* 27(1): 43-47.

Stallings, Jane A., Stephanie L. Knight, and David Markham. 2014. "Using the Stallings Observation System to investigate time on task in four countries." World Bank Report No. 92558. http://documents.worldbank.org/curated/en/496851468182672630/Using-the-stallings-observation-system-to-investigate-time-on-task-in-four-countries

Wolf, Sharon, Mahjabeen Raza, Sharon Kim, J. Lawrence Aber, Jere Behrman, and Edward Seidman. 2018. "Measuring and predicting process quality in Ghanaian pre-primary classrooms using the Teacher Instructional Practices and Processes System (TIPPS). *Early Childhood Research Quarterly 45*: 18-30.

World Bank. 2004. World Development Report 2003: Making Services Work for Poor People. Washington, DC: Oxford University Press and The World Bank.

World Bank. 2015. *Conducting Classroom Observations: Analyzing Classrooms Dynamics and Instructional Time*. Washington, DC.

World Bank. 2018. World Development Report 2018: Learning to Realize Education's Promise. Washington, DC: Oxford University Press and The World Bank.

**Figure 1: Coefficients of variation for Level 1 variables from the different observation instruments**

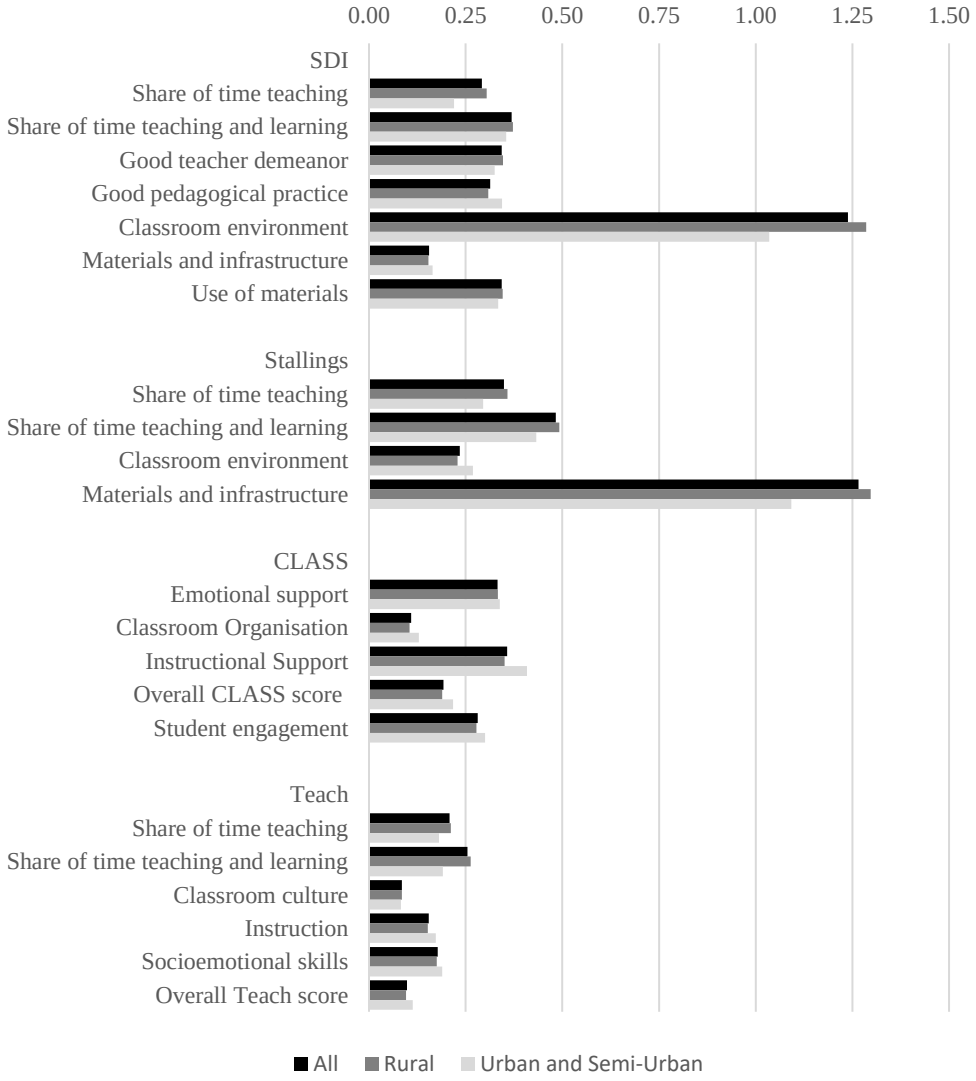
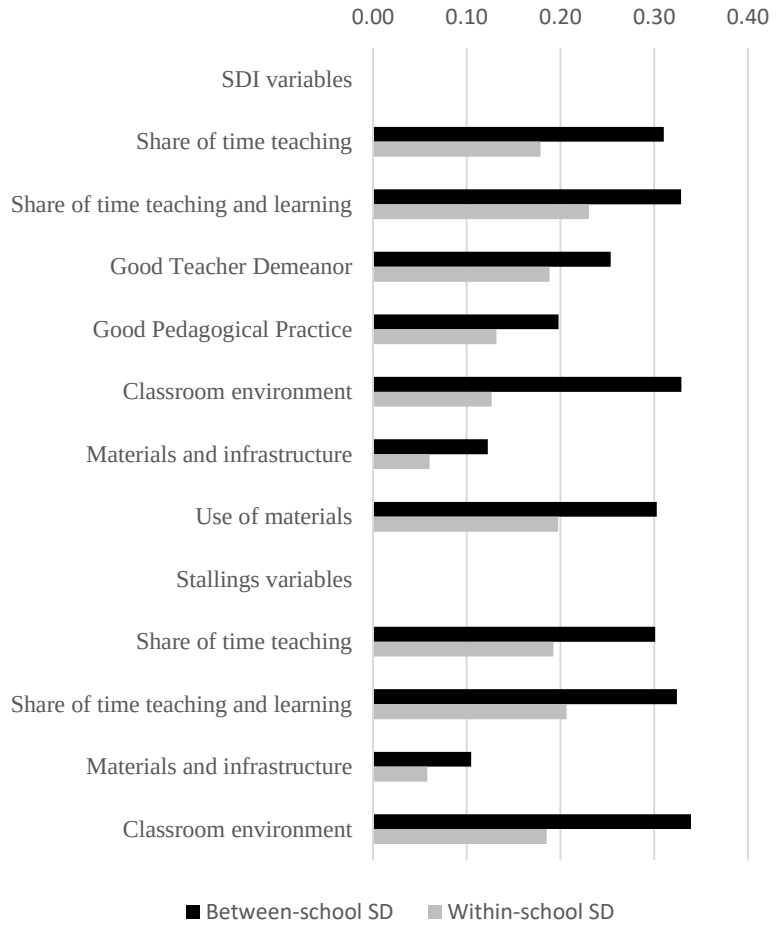


**Figure 2: Between-school and within-school variation in SDI and Stallings Level 1 variables**




Note: Between-school variation is the standard deviation of Grade 4 values across schools. Within-school variation is the mean, across schools, of the within-school standard deviation of each variable across grades 3, 4, and, 5

**Figure 3: Distribution of share of time variables**

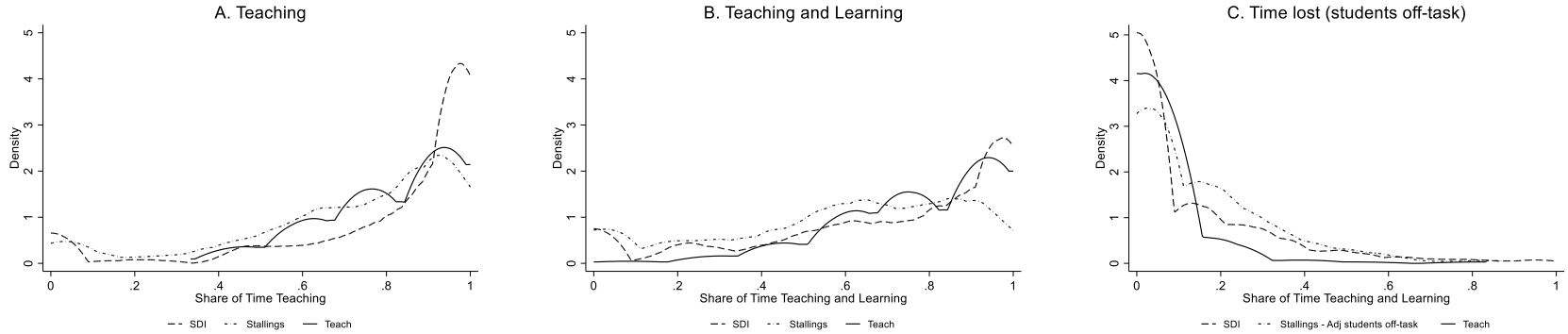


**Table 1A: Summary Statistics from schools in ETOS sample**

| | All (N=106) | | | | Rural (N=88) | | | | Urban (N=18) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | SD | 25 pctile | 75th pctile | Mean | SD | 25 pctile | 75th pctile | Mean | SD | 25 pctile | 75th pctile |
| Number of teachers | 14 | 11 | 7 | 15 | 11 | (6) | 7 | 12 | 30 | (16) | 16 | 47 |
| Total Enrollment (G1=G6) | 594 | 500 | 326 | 638 | 494 | (293) | 319 | 606 | 1084 | (895) | 386 | 1505 |
| Average class size | 58 | 30 | 38 | 75 | 54 | (25) | 37 | 64 | 77 | (44) | 40 | 116 |
| Average teacher absence from school | 0.14 | 0.15 | 0.00 | 0.20 | 0.13 | (0.13) | 0.00 | 0.20 | 0.18 | (0.20) | 0.00 | 0.29 |
| Average teacher absence from classroom | 0.39 | 0.20 | 0.29 | 0.50 | 0.38 | (0.19) | 0.25 | 0.50 | 0.42 | (0.25) | 0.30 | 0.50 |

Source: SDI 2016 data for ETOS schools

**Table 1B: Summary Statistics from schools in nationally representative SDI sample**

| | All (N=400) | | | | Rural (N=271) | | | | Urban (N=129) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | SD | 25 pctile | 75th pctile | Mean | SD | 25 pctile | 75th pctile | Mean | SD | 25 pctile | 75th pctile |
| Number of teachers | 17 | (12) | 9 | 22 | 12 | (6) | 8 | 14 | 29 | (14) | 18 | 42 |
| Total Enrollment (G1=G6) | 746 | (601) | 400 | 853 | 577 | (335) | 361 | 686 | 1102 | (837) | 537 | 1353 |
| Average class size | 70 | (46) | 39 | 86 | 62 | (37) | 38 | 76 | 88 | (56) | 46 | 115 |
| Average teacher absence from school | 0.14 | (0.16) | 0.00 | 0.20 | 0.14 | (0.15) | 0.00 | 0.20 | 0.14 | (0.16) | 0.00 | 0.20 |
| Average teacher absence from classroom | 0.42 | (0.22) | 0.29 | 0.50 | 0.41 | (0.22) | 0.25 | 0.56 | 0.42 | (0.24) | 0.30 | 0.50 |

Source: SDI 2016 data

**Table 2: Summary statistics on classroom conditions during ETOS observations**

| | All (N=176) | | Rural (N=148) | | Urban (N=28) | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD |
| Blackboard in class (1=yes) | 0.99 | (0.11) | 0.99 | (0.12) | 1.00 | (0.00) |
| Blackboard has sufficient contrast for reading (1=yes) | 0.90 | (0.30) | 0.91 | (0.29) | 0.86 | (0.36) |
| Chalk available during the lesson (1=yes) | 0.98 | (0.13) | 0.98 | (0.14) | 1.00 | (0.00) |
| Classroom has a working electricity connection (1=yes) | 0.13 | (0.34) | 0.10 | (0.30) | 0.29 | (0.46) |
| Hygiene in class (1 = reasonably or extremely clean) | 0.76 | (0.43) | 0.75 | (0.43) | 0.79 | (0.42) |
| Pupils work displayed on the the wall | 0.25 | (0.43) | 0.23 | (0.42) | 0.36 | (0.49) |
| Charts displayed on the walls | 0.36 | (0.48) | 0.34 | (0.48) | 0.43 | (0.50) |
| Corner Library | 0.04 | (0.20) | 0.05 | (0.21) | 0.00 | (0.00) |

**Table 3: Summary statistics on selected teacher practices observed during classroom observations**

| | All (N=176) | | Rural (N=148) | | Urban (N=28) | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD |
| **Use of materials** | | | | | | |
| Textbook was used by teacher | 0.84 | (0.37) | 0.83 | (0.38) | 0.86 | (0.36) |
| Teacher writes on the blackboard | 0.94 | (0.23) | 0.94 | (0.24) | 0.96 | (0.19) |
| Pupils write on the blackboard | 0.52 | (0.50) | 0.51 | (0.50) | 0.61 | (0.50) |
| Teacher uses local information to make learning relevant | 0.54 | (0.50) | 0.55 | (0.50) | 0.46 | (0.51) |
| **Teacher demeanor** | | | | | | |
| Teacher mostly standing (as opposed to sitting) | 0.88 | (0.33) | 0.88 | (0.33) | 0.86 | (0.36) |
| Teacher visits children individually | 0.61 | (0.49) | 0.61 | (0.49) | 0.57 | (0.50) |
| Teacher calls pupil by name while teaching | 0.89 | (0.31) | 0.89 | (0.31) | 0.89 | (0.31) |
| Teacher smiling, laughing, or joking with pupils | 0.53 | (0.50) | 0.55 | (0.50) | 0.43 | (0.50) |
| Teacher hit pupil (1=no) | 0.85 | (0.36) | 0.84 | (0.36) | 0.89 | (0.31) |
| **Pedagogical practices** | | | | | | |
| Teacher ask questions that required learners to recall information | 0.68 | (0.47) | 0.69 | (0.46) | 0.64 | (0.49) |
| Teacher ask learners to to demonstrate their understanding | 0.89 | (0.32) | 0.89 | (0.31) | 0.86 | (0.36) |
| Teacher ask questions that required learners to apply information | 0.79 | (0.41) | 0.78 | (0.41) | 0.82 | (0.39) |
| Teacher ask questions requiring learners to use their creativity | 0.83 | (0.38) | 0.83 | (0.38) | 0.82 | (0.39) |
| Teacher gives feedback and/or encouragement to students (scale, 1=yes) | 0.63 | (0.44) | 0.63 | (0.44) | 0.67 | (0.42) |
| Teacher gives feedback that was correcting a mistake (scale, 1=yes) | 0.68 | (0.43) | 0.69 | (0.42) | 0.62 | (0.48) |
| Teachers gives feedback that was scolding a mistake (scale, 1=no) | 0.79 | (0.37) | 0.78 | (0.38) | 0.86 | (0.32) |
| Teacher introduces the lesson at the start of class | 0.89 | (0.32) | 0.89 | (0.32) | 0.89 | (0.31) |
| Teacher summarize the lesson at the end of class | 0.21 | (0.41) | 0.20 | (0.40) | 0.25 | (0.44) |
| Teacher assigns homework | 0.16 | (0.37) | 0.18 | (0.38) | 0.11 | (0.31) |
| Teacher reviews homeworks | 0.13 | (0.33) | 0.14 | (0.34) | 0.07 | (0.26) |
| Teacher uses local langague for instructions | 0.06 | (0.24) | 0.05 | (0.21) | 0.14 | (0.36) |

Notes: All answers are binary 1=yes 0=no, unless otherwise indicated as scale variables (where 3 categories were rescaled to be between 0 and 1)

**Table 4: Summary Statistics derived from classroom observation instruments (Level 1 variables)**

| | All | | | | Rural | | | | Urban and Semi-Urban | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | SD | 25 ptile | 75 ptile | Mean | SD | 25 ptile | 75 ptile | Mean | SD | 25 ptile | 75 ptile |
| **SDI** | | | | | | | | | | | | |
| Share of time teaching | 0.85 | (0.25) | 0.81 | 1.00 | 0.8378 | (0.25) | 0.79 | 1.00 | 0.90 | (0.20) | 0.91 | 0.99 |
| Share of time teaching and learni | 0.75 | (0.28) | 0.61 | 0.98 | 0.74 | (0.28) | 0.60 | 0.98 | 0.78 | (0.28) | 0.62 | 0.98 |
| Good teacher demeanor | 0.67 | (0.23) | 0.57 | 0.86 | 0.68 | (0.23) | 0.57 | 0.86 | 0.63 | (0.20) | 0.57 | 0.75 |
| Good pedagogical practice | 0.56 | (0.18) | 0.49 | 0.69 | 0.56 | (0.17) | 0.49 | 0.69 | 0.55 | (0.19) | 0.48 | 0.68 |
| Classroom environment | 0.30 | (0.37) | 0.00 | 0.50 | 0.28 | (0.36) | 0.00 | 0.50 | 0.40 | (0.42) | 0.00 | 0.88 |
| Materials and infrastructure | 0.68 | (0.11) | 0.57 | 0.71 | 0.68 | (0.10) | 0.57 | 0.71 | 0.70 | (0.11) | 0.57 | 0.71 |
| Use of materials | 0.76 | (0.26) | 0.67 | 1.00 | 0.76 | (0.26) | 0.67 | 1.00 | 0.80 | (0.27) | 0.67 | 1.00 |
| **Stallings** | | | | | | | | | | | | |
| Share of time teaching | 0.73 | (0.25) | 0.60 | 0.93 | 0.72 | (0.26) | 0.60 | 0.92 | 0.78 | (0.23) | 0.63 | 0.98 |
| Share of time teaching and learni | 0.61 | (0.29) | 0.41 | 0.87 | 0.60 | (0.30) | 0.40 | 0.87 | 0.65 | (0.28) | 0.50 | 0.86 |
| Classroom environment | 0.28 | (0.10) | 0.40 | 0.47 | 0.42 | (0.10) | 0.40 | 0.47 | 0.44 | (0.12) | 0.40 | 0.53 |
| Materials and infrastructure | 0.43 | (0.35) | 0.00 | 0.50 | 0.28 | (0.36) | 0.00 | 0.50 | 0.29 | (0.32) | 0.00 | 0.50 |
| **CLASS (7-point scales)** | | | | | | | | | | | | |
| Emotional support | 2.92 | (0.97) | 2.33 | 3.67 | 2.93 | (0.98) | 2.33 | 3.67 | 2.84 | (0.96) | 2.08 | 3.50 |
| Classroom organisation | 5.75 | (0.63) | 5.42 | 6.17 | 5.78 | (0.61) | 5.50 | 6.17 | 5.58 | (0.72) | 5.33 | 6.08 |
| Instructional support | 2.66 | (0.95) | 2.00 | 3.20 | 2.67 | (0.94) | 2.00 | 3.18 | 2.58 | (1.05) | 1.80 | 3.38 |
| Overall CLASS score | 3.78 | (0.73) | 3.27 | 4.24 | 3.79 | (0.72) | 3.29 | 4.26 | 3.67 | (0.80) | 3.13 | 4.14 |
| Student engagement | 4.02 | (1.13) | 3.50 | 5.00 | 4.06 | (1.13) | 3.50 | 5.00 | 3.80 | (1.14) | 3.00 | 4.75 |
| **Teach (5-point scales)** | | | | | | | | | | | | |
| Share of time teaching | 0.84 | (0.17) | 0.67 | 1.00 | 0.83 | (0.18) | 0.67 | 1.00 | 0.89 | (0.16) | 0.83 | 1.00 |
| Share of time teaching and learni | 0.81 | (0.21) | 0.67 | 1.00 | 0.80 | (0.21) | 0.67 | 1.00 | 0.88 | (0.17) | 0.83 | 1.00 |
| Classroom culture | 3.65 | (0.31) | 3.50 | 4.00 | 3.64 | (0.31) | 3.50 | 4.00 | 3.68 | (0.30) | 3.50 | 4.00 |
| Instruction | 2.43 | (0.38) | 2.25 | 2.75 | 2.44 | (0.37) | 2.25 | 2.75 | 2.40 | (0.41) | 2.19 | 2.75 |
| Socioemotional skills | 2.07 | (0.37) | 1.83 | 2.33 | 2.06 | (0.36) | 1.83 | 2.17 | 2.15 | (0.41) | 1.83 | 2.33 |
| Overall Teach score | 2.46 | (0.24) | 2.30 | 2.64 | 2.46 | (0.24) | 2.28 | 2.61 | 2.51 | (0.28) | 2.33 | 2.74 |

**Table 5: Inter-Rater Reliability (ICC using one-way random effects model)**

| Level 1 | | | Level 2 | | |
|---|---|---|---|---|---|
| | ICC | Mean diff. (abs. val.) | | ICC | Mean diff. (abs. val.) |
| **SDI** | | | | | |
| Share of time teaching | **0.95** | 0.04 | Teacher interacts with students as a group | **0.86** | 0.08 |
| Share of time teaching and learning | **0.75** | 0.15 | Teacher interacts with a small group of children | 0.00 | 0.04 |
| Good teacher demeanor | **0.87** | 0.09 | Teacher interacts with children one on one | **0.33** | 0.10 |
| Good pedagogical practice | **0.73** | 0.07 | Teacher reads, lectures, or demonstrates to the pupils | 0.00 | 0.03 |
| Classroom environment | **0.64** | 0.23 | Teacher supervises pupils writing on the board | **0.95** | 0.02 |
| Materials and infrastrucutre | **0.46** | 0.07 | Teacher leads kinesthetic group learning activity | **0.42** | 0.02 |
| Use of materials | **0.84** | 0.10 | Teacher writing on blackboard | **0.71** | 0.06 |
| | | | Teacher listening to pupils read/recite | 0.00 | 0.03 |
| | | | Teacher waiting for pupils to complete task | **0.72** | 0.07 |
| | | | Teacher testing students in class | 0.00 | 0.04 |
| | | | Teacher maintaining discipline in class | 0.00 | 0.00 |
| | | | Teacher in class, not teaching | 0.00 | 0.01 |
| | | | Teacher not in class, learning activity ongoing | **0.93** | 0.01 |
| | | | Teacher not in class, no learning activity ongoing | **0.96** | 0.03 |
| | | | Good Teacher Demeanor | **0.87** | 0.09 |
| | | | Good Pedagogical Practise | **0.73** | 0.07 |
| | | | Classroom Environment | **0.64** | 0.23 |
| | | | Availability of materials | **0.46** | 0.07 |
| | | | Use of materials | **0.84** | 0.10 |
| **Stallings** | | | | | |
| Share of time teaching | **0.77** | 0.08 | Teacher involved in reading activity | **0.79** | 0.01 |
| Share of time teaching and learning | **0.63** | 0.17 | Teacher involved in lecture activity | **0.92** | 0.04 |
| Materials and infrastrucutre | **0.52** | 0.04 | Teacher involved in discussion | **0.79** | 0.07 |
| Classroom environment | **0.63** | 0.21 | Teacher involved in practice activity | 0.00 | 0.02 |
| | | | Teacher involved in class work activity | **0.77** | 0.07 |
| | | | Teacher giving instructions on blackboard | **0.86** | 0.01 |
| | | | Teacher giving verbal instructions to students | **0.75** | 0.01 |
| | | | Teacher involved in classroom management related act | **0.49** | 0.01 |
| | | | Teacher involved in classroom management related act | **0.79** | 0.02 |
| | | | Teacher not present in the classroom | **0.84** | 0.04 |
| | | | Materials and infrastrucutre | **0.52** | 0.04 |
| | | | Classroom environment | **0.63** | 0.21 |

continued...

**Table 5 continued: Inter-Rater Reliability (ICC using one-way random effects model)**

| Level 1 | | | Level 2 | | |
|---|---|---|---|---|---|
| | ICC | Mean diff. (abs. val.) | | ICC | Mean diff. (abs. val.) |
| **CLASS** | | | | | |
| Emotional support | 0.00 | 1.30 | Positive Climate | 0.06 | 1.52 |
| Classroom organisation | 0.20 | 0.57 | Teacher Sensitivity | 0.00 | 1.77 |
| Instructional support | 0.00 | 1.19 | Regard for Adolescent Perspective | 0.05 | 0.91 |
| Overall CLASS score | 0.00 | 0.85 | Behavior Management | 0.00 | 1.02 |
| Student engagement | 0.09 | 1.25 | Teacher Productivity | 0.17 | 1.36 |
| | | | Negative Climate | 0.00 | 0.59 |
| | | | Instructional Learning Format | 0.00 | 1.43 |
| | | | Content Understanding | 0.05 | 1.23 |
| | | | Analysis and Inquiry | 0.00 | 1.43 |
| | | | Quality of Feedback | 0.00 | 1.27 |
| | | | Instructional Dialogue | 0.00 | 1.36 |
| | | | Student Engagement | 0.09 | 1.25 |
| **Teach** | | | | | |
| Share of time teaching | **0.90** | 0.02 | Supportive Learning Environment | **0.56** | 0.31 |
| Share of time teaching and learning | **0.74** | 0.04 | Positive Behavioral Expectations | 0.00 | 0.32 |
| Classroom culture | **0.57** | 0.17 | Lesson facilitation | **0.73** | 0.38 |
| Instruction | **0.51** | 0.29 | Checks for understanding | **0.59** | 0.53 |
| Socioemotional skills | **0.84** | 0.15 | Feedback | **0.52** | 0.36 |
| Overall Teach score | **0.76** | 0.13 | Critical Thinking | 0.30 | 0.18 |
| | | | Autonomy | **0.78** | 0.25 |
| | | | Perseverance | 0.01 | 0.04 |
| | | | Social and collaborative skills | **0.86** | 0.24 |

Note: Number of lessons/videos double-coded: SDI 27; Stallings 28; CLASS 44; Teach 55. Values are truncated at 0. ICCs greater than 0.3 are in bold.

**Table 6: Inter-rater reliability (Percent of raters scoring within 1 or 2 points of each other)**

| | Level 1 | | | Level 2 | |
|---|---|---|---|---|---|
| | Within 1 point | Within 2 points | | Within 1 point | Within 2 points |
| **CLASS (7-point scale)** | | | | | |
| Emotional support | 43.1 | 81.8 | Positive Climate | 45.5 | 88.6 |
| Classroom organisation | 95.5 | 100 | Teacher Sensitivity | 47.7 | 77.3 |
| Instructional support | 45.5 | 84.1 | Regard for Adolescent Perspective | 79.5 | 95.5 |
| Overall CLASS score | 56.8 | 97.7 | Behavior Management | 79.5 | 93.2 |
| Student engagement | 61.4 | 90.1 | Teacher Productivity | 63.6 | 75 |
| | | | Negative Climate | 97.7 | 100 |
| | | | Instructional Learning Format | 56.8 | 84.1 |
| | | | Content Understanding | 65.9 | 84.1 |
| | | | Analysis and Inquiry | 54.5 | 81.8 |
| | | | Quality of Feedback | 65.9 | 86.4 |
| | | | Instructional Dialogue | 59.1 | 84.1 |
| | | | Student Engagement | 61.4 | 90.1 |
| **Teach (5-point scale)** | | | | | |
| Classroom culture | 100 | 100 | Supportive Learning Environment | 96.4 | 100 |
| Instruction | 98.2 | 100 | Positive Behavioral Expectations | 100 | 100 |
| Socioemotional skills | 100 | 100 | Lesson facilitation | 98.2 | 100 |
| Overall Teach score | 100 | 100 | Checks for understanding | 96.4 | 100 |
| | | | Feedback | 96.4 | 100 |
| | | | Critical Thinking | 96.4 | 100 |
| | | | Autonomy | 100 | 100 |
| | | | Perseverance | 100 | 100 |
| | | | Social and collaborative skills | 100 | 100 |

**Table 7: Correlations between Level 1 variables within each tool**

| **SDI (N=268)** | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| (1) Share of time teaching | **1.00** | | | | | | |
| (2) Share of time teaching and learnir | **0.81** | **1.00** | | | | | |
| (3) Good Teacher Demeanor | **0.66** | **0.65** | **1.00** | | | | |
| (4) Good Pedagogical Practise | **0.63** | **0.58** | **0.68** | **1.00** | | | |
| (5) Classroom environment | 0.13 | 0.11 | 0.08 | -0.02 | **1.00** | | |
| (6) Materials and infrastructure | 0.11 | 0.15 | 0.08 | 0.06 | 0.20 | **1.00** | |
| (7) Use of materials | **0.64** | **0.52** | **0.63** | **0.61** | 0.01 | 0.00 | **1.00** |

| **Stallings (N=277)** | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| (1) Share of time teaching | **1.00** | | | |
| (2) Share of time teaching and learnir | **0.83** | **1.00** | | |
| (3) Materials and infrastructure | **0.34** | **0.31** | **1.00** | |
| (4) Classroom environment | -0.03 | -0.01 | 0.07 | **1.00** |

| **CLASS (N=149)** | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| (1) Emotional Support | **1.00** | | | | |
| (2) Classroom Organisation | **0.49** | **1.00** | | | |
| (3) Instructional Support | **0.81** | **0.37** | **1.00** | | |
| (4) Overall CLASS score | **0.94** | **0.67** | **0.90** | **1.00** | |
| (5) Student Engagement | **0.80** | **0.55** | **0.79** | **0.85** | **1.00** |

| **Teach (N=151)** | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| (1) Share of time teaching | **1.00** | | | | | |
| (2) Share of time teaching and learnir | **0.88** | **1.00** | | | | |
| (3) Classroom culture | 0.06 | 0.08 | **1.00** | | | |
| (4) Instruction | 0.13 | 0.12 | 0.15 | **1.00** | | |
| (5) Socioemotional Skills | -0.03 | 0.01 | **0.34** | 0.19 | **1.00** | |
| (6) Overall Teach score | 0.07 | 0.09 | **0.60** | **0.68** | **0.75** | **1.00** |

Correlation coefficients greater than 0.3 are in bold.

**Table 8: Principal Components Analysis of Level 1 variables**

| | Summary Statistics | | | | Component loadings for first three principal components | | |
|---|---|---|---|---|---|---|---|
| | Eigen-value | Diff-erence | Prop. | | Comp1 | Comp2 | Comp3 |
| **SDI (N=268)** | | | | | | | |
| Comp 1 | **2.86** | 1.65 | 0.48 | Share of time teaching and learning | **0.49** | 0.09 | -0.02 |
| Comp 2 | **1.21** | 0.40 | 0.20 | Good Teacher Demeanor | **0.52** | -0.02 | 0.04 |
| Comp 3 | 0.80 | 0.35 | 0.13 | Good Pedagogical Practise | **0.50** | -0.12 | -0.07 |
| | | | | Classroom environment | 0.06 | **0.70** | **0.70** |
| | | | | Materials and infrastructure | 0.08 | **0.69** | **-0.70** |
| | | | | Use of materials | **0.48** | -0.15 | 0.08 |
| **Stallings (N=277)** | | | | | | | |
| Comp 1 | **1.32** | 0.31 | 0.44 | Share of time teaching and learning | **0.69** | -0.22 | **0.69** |
| Comp 2 | **1.00** | 0.32 | 0.33 | Materials and infrastructure | **0.71** | 0.02 | **-0.71** |
| Comp 3 | 0.68 | . | 0.23 | Classroom environment | 0.14 | **0.98** | 0.17 |
| **CLASS (N=148)** | | | | | | | |
| Comp 1 | **2.94** | 2.25 | 0.74 | Emotional Support | **0.54** | -0.20 | **0.62** |
| Comp 2 | 0.69 | 0.49 | 0.17 | Classroom Organisation | **0.39** | **0.89** | 0.12 |
| Comp 3 | 0.20 | 0.03 | 0.05 | Instructional Support | **0.52** | -0.41 | 0.08 |
| | | | | Student Engagement | **0.54** | -0.05 | **-0.77** |
| **Teach (N=151)** | | | | | | | |
| Comp 1 | **1.48** | 0.46 | 0.37 | Share of time teaching and learning | 0.23 | **0.85** | **0.42** |
| Comp 2 | **1.02** | 0.17 | 0.26 | Classroom culture | **0.60** | -0.21 | **0.41** |
| Comp 3 | 0.85 | 0.20 | 0.21 | Instruction | **0.47** | **0.31** | **-0.81** |
| | | | | Socioemotional Skills | **0.60** | **-0.36** | 0.06 |

Note: Eigen values greater than 1 and component loadings greater than 0.3 (in absolute value) are in bold.

**Table 9: Principal Components Analysis of Level 2 variables**

| | Summary Statistics | | | Component loadings for first five principal components | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Eigen- | Diff- | Prop. | Variable | Comp1 | Comp2 | Comp3 | Comp4 | Comp5 |
| **SDI** | | | | | | | | | |
| Comp 1 | 3.76 | 2.02 | 0.18 | Teacher interacts with students as a group | 0.21 | 0.28 | 0.24 | -0.17 | -0.24 |
| Comp 2 | 1.74 | 0.27 | 0.08 | Teacher interacts with a small group of children | 0.05 | 0.12 | 0.00 | -0.32 | 0.21 |
| Comp 3 | 1.47 | 0.10 | 0.07 | Teacher interacts with children one on one | 0.14 | 0.19 | -0.37 | 0.23 | 0.02 |
| Comp 4 | 1.36 | 0.12 | 0.06 | Teacher reads, lectures, or demonstrates to the pupils | 0.12 | -0.46 | 0.11 | -0.04 | 0.03 |
| Comp 5 | 1.24 | 0.10 | 0.06 | Teacher supervises pupils writing on the board | 0.15 | -0.13 | -0.07 | -0.18 | -0.62 |
| | | | | Teacher leads kinesthetic group learning activity | 0.05 | -0.04 | 0.05 | -0.22 | 0.09 |
| | | | | Teacher writing on blackboard | 0.26 | -0.04 | -0.04 | **0.31** | -0.09 |
| | | | | Teacher listening to pupils read/recite | 0.05 | -0.12 | 0.30 | -0.47 | **0.33** |
| | | | | Teacher waiting for pupils to complete task | 0.17 | 0.07 | -0.37 | 0.00 | **0.40** |
| | | | | Teacher testing students in class | 0.03 | **0.34** | **0.35** | 0.20 | -0.11 |
| | | | | Teacher maintaining discipline in class | 0.08 | 0.21 | **0.40** | 0.09 | -0.04 |
| | | | | Teacher doing paperwork | -0.01 | 0.04 | 0.02 | 0.11 | 0.21 |
| | | | | Teacher in class, not teaching | 0.05 | 0.18 | **0.36** | 0.23 | 0.14 |
| | | | | Teacher not in class, learning activity ongoing | 0.04 | -0.30 | 0.12 | **0.37** | 0.21 |
| | | | | Teacher not in class, no learning activity ongoing | -0.48 | -0.07 | -0.03 | -0.02 | -0.07 |
| | | | | Break | -0.01 | -0.11 | 0.21 | **0.30** | 0.12 |
| | | | | Good Teacher Demeanor | **0.43** | -0.01 | -0.03 | -0.01 | 0.13 |
| | | | | Good Pedagogical Practise | **0.43** | -0.14 | 0.05 | 0.00 | 0.07 |
| | | | | Classroom Environment | 0.05 | **0.41** | -0.02 | -0.25 | 0.18 |
| | | | | Materials and Infrastructure | 0.06 | **0.35** | -0.28 | 0.11 | -0.04 |
| | | | | Use of materials | **0.42** | -0.11 | -0.03 | -0.07 | -0.17 |
| **Stallings** | | | | | | | | | |
| Comp 1 | 2.67 | 0.67 | 0.18 | Teacher involved in reading activity | -0.56 | 0.13 | 0.01 | 0.03 | 0.05 |
| Comp 2 | 1.99 | 0.68 | 0.13 | Teacher involved in lecture activity | 0.00 | -0.18 | **0.46** | -0.48 | 0.18 |
| Comp 3 | 1.32 | 0.10 | 0.09 | Teacher involved in discussion | 0.19 | 0.00 | 0.19 | **0.72** | -0.21 |
| Comp 4 | 1.22 | 0.08 | 0.08 | Teacher involved in practice activity | 0.21 | 0.20 | -0.45 | 0.05 | 0.15 |
| Comp 5 | 1.14 | 0.01 | 0.08 | Teacher involved in class work activity | 0.25 | -0.41 | -0.36 | -0.25 | -0.01 |
| | | | | Teacher giving instructions on blackboard | 0.12 | 0.04 | **0.33** | -0.23 | -0.49 |
| | | | | Teacher giving verbal instructions to students | 0.15 | **0.34** | 0.12 | -0.10 | -0.21 |
| | | | | Teacher involved in social interactions | 0.19 | **0.43** | 0.06 | -0.11 | 0.10 |
| | | | | Teacher involved in dicipline related activity | 0.17 | **0.44** | 0.12 | -0.18 | 0.12 |
| | | | | Teacher involved in classroom management related activit | 0.19 | **0.37** | 0.03 | 0.04 | 0.26 |
| | | | | Teacher involved in classroom management related activit | -0.09 | -0.15 | **0.31** | 0.16 | **0.62** |
| | | | | Teacher uninvolved | -0.02 | 0.08 | -0.04 | 0.03 | **0.33** |
| | | | | Teacher not present in the classroom | -0.54 | 0.20 | -0.14 | 0.01 | -0.11 |
| | | | | Materials and Infrastructure | **0.33** | -0.18 | 0.05 | 0.13 | 0.15 |
| | | | | Classroom environment | -0.03 | -0.08 | **0.40** | 0.18 | -0.07 |

continued…

**Table 9 continued: Principal Components Analysis of Level 2 Variables**

| | Summary Statistics | | | | Coefficients for first five principal components | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Eigen- | Diff- | Prop. | Variable | Comp1 | Comp2 | Comp3 | Comp4 | Comp5 |
| **CLASS** | | | | | | | | | |
| Comp 1 | 7.06 | 5.58 | 0.59 | Positive Climate | **0.31** | 0.19 | 0.07 | -0.32 | -0.03 |
| Comp 2 | 1.47 | 0.54 | 0.12 | Teacher Sensitivity | **0.31** | 0.19 | 0.14 | -0.13 | -0.61 |
| Comp 3 | 0.93 | 0.21 | 0.08 | Regard for Adolescent Perspective | **0.31** | -0.12 | 0.25 | -0.01 | -0.39 |
| Comp 4 | 0.72 | 0.33 | 0.06 | Behavior Management | 0.00 | **0.62** | -0.37 | **0.63** | -0.17 |
| Comp 5 | 0.39 | 0.03 | 0.03 | Teacher Productivity | 0.29 | 0.29 | -0.08 | -0.26 | **0.50** |
| | | | | Negative Climate | -0.10 | **0.33** | **0.87** | 0.26 | 0.22 |
| | | | | Instructional Learning Format | **0.34** | 0.17 | -0.02 | 0.00 | 0.02 |
| | | | | Content Understanding | **0.33** | 0.09 | -0.06 | -0.06 | 0.11 |
| | | | | Analysis and Inquiry | 0.27 | -0.38 | 0.01 | **0.46** | 0.26 |
| | | | | Quality of Feedback | **0.33** | -0.24 | 0.00 | 0.19 | 0.03 |
| | | | | Instructional Dialogue | **0.31** | -0.28 | 0.07 | 0.30 | -0.06 |
| | | | | Student Engagement | **0.33** | 0.13 | -0.10 | -0.05 | 0.27 |
| **Teach** | | | | | | | | | |
| Comp 1 | 2.04 | 0.56 | 0.20 | Share of time teaching and learning | 0.11 | 0.13 | **0.31** | **0.59** | 0.20 |
| Comp 2 | 1.48 | 0.17 | 0.15 | Supportive Learning Environment | **0.46** | -0.39 | -0.03 | 0.00 | **0.31** |
| Comp 3 | 1.31 | 0.22 | 0.13 | Positive Behavioral Expectations | -0.13 | 0.06 | **0.31** | **0.64** | -0.12 |
| Comp 4 | 1.09 | 0.14 | 0.11 | Lesson facilitation | **0.36** | -0.17 | **0.52** | -0.09 | 0.02 |
| Comp 5 | 0.95 | 0.09 | 0.10 | Checks for understanding | **0.38** | **0.39** | -0.12 | 0.01 | -0.44 |
| | | | | Feedback | 0.27 | **0.51** | -0.16 | 0.07 | -0.07 |
| | | | | Critical Thinking | 0.28 | **0.43** | 0.26 | -0.26 | 0.09 |
| | | | | Autonomy | **0.52** | -0.30 | 0.04 | -0.01 | -0.12 |
| | | | | Perseverance | 0.19 | 0.17 | -0.47 | 0.22 | **0.66** |
| | | | | Social and collaborative skills | 0.18 | -0.28 | -0.46 | **0.34** | -0.45 |

**Table 10: Correlation of Level 1 variables across tools**

**A. SDI vs. Stallings (N=153)**

| | | Stallings | | | |
|---|---|---|---|---|---|
| | | Share of time teaching | Share of time teaching and learning | Materials and infrastructure | Classroom environment |
| SDI | Share of time teaching | **0.80** | **0.66** | 0.29 | 0.00 |
| | Share of time teaching and learning | **0.66** | **0.70** | 0.25 | 0.04 |
| | Good Teacher Demeanor | **0.36** | **0.31** | 0.22 | 0.07 |
| | Good Pedagogical Practise | **0.33** | 0.25 | 0.19 | -0.03 |
| | Classroom environment | 0.06 | 0.06 | -0.02 | **0.69** |
| | Materials and infrastructure | 0.20 | 0.24 | 0.21 | 0.04 |
| | Use of materials | **0.39** | 0.27 | 0.21 | 0.00 |

**B. SDI vs. CLASS (N=129)**

| | | CLASS | | | | |
|---|---|---|---|---|---|---|
| | | Emotional Support | Classroom Organisation | Instructional Support | Overall CLASS | Student Engagement |
| SDI | Share of time teaching | 0.23 | **0.55** | 0.24 | **0.37** | 0.31 |
| | Share of time teaching and learning | 0.19 | **0.47** | 0.26 | **0.34** | 0.29 |
| | Good Teacher Demeanor | 0.20 | 0.24 | 0.20 | 0.24 | 0.19 |
| | Good Pedagogical Practise | 0.18 | 0.23 | 0.20 | 0.24 | 0.28 |
| | Classroom environment | 0.05 | 0.02 | 0.02 | 0.03 | 0.01 |
| | Materials and infrastructure | -0.04 | 0.17 | -0.03 | 0.02 | -0.05 |
| | Use of materials | 0.16 | 0.19 | 0.14 | 0.18 | 0.15 |

**C. SDI vs. Teach (N=130)**

| | | Teach | | | | | |
|---|---|---|---|---|---|---|---|
| | | Share of time teaching | Share of time teaching and learning | Classroom culture | Instruction | Socio-emotional Skills | Overall Teach |
| SDI | Share of time teaching | **0.50** | **0.39** | -0.02 | 0.19 | 0.00 | 0.09 |
| | Share of time teaching and learning | **0.33** | **0.32** | 0.04 | 0.14 | -0.02 | 0.07 |
| | Good Teacher Demeanor | 0.09 | 0.11 | 0.09 | 0.14 | 0.10 | 0.15 |
| | Good Pedagogical Practise | 0.07 | 0.08 | 0.12 | 0.18 | 0.08 | 0.21 |
| | Classroom environment | -0.08 | -0.05 | 0.02 | 0.01 | 0.06 | 0.04 |
| | Materials and infrastructure | 0.03 | 0.00 | 0.07 | 0.12 | -0.10 | 0.05 |
| | Use of materials | 0.09 | 0.09 | 0.06 | 0.30 | 0.23 | 0.29 |

continued…

**Table 10 continued: Correlation of Level 1 variables across tools**

**D. Stallings vs. CLASS (N=127)**

| | | CLASS | | | | |
|---|---|---|---|---|---|---|
| | | Emotional Support | Classroom Organisation | Instructional Support | Overall CLASS | Student Engagement |
| **Stallings** | Share of time teaching | 0.22 | **0.43** | 0.19 | **0.30** | 0.18 |
| | Share of time teaching and learning | 0.30 | **0.42** | 0.28 | **0.37** | 0.23 |
| | Materials and infrastructure | 0.14 | 0.27 | 0.13 | 0.20 | 0.15 |
| | Classroom environment | 0.13 | 0.12 | 0.09 | 0.13 | 0.11 |

**E. Stallings vs. Teach (N=132)**

| | | Teach | | | | | |
|---|---|---|---|---|---|---|---|
| | | Share of time teaching | Share of time teaching and learning | Classroom culture | Instruction | Socioemotional Skills | Overall Teach |
| **Stallings** | Share of time teaching | **0.43** | **0.33** | -0.15 | 0.21 | -0.17 | -0.02 |
| | Share of time teaching and learning | **0.41** | **0.41** | 0.00 | 0.22 | -0.10 | 0.06 |
| | Materials and infrastructure | 0.19 | 0.22 | 0.26 | 0.05 | 0.04 | 0.14 |
| | Classroom environment | -0.04 | 0.00 | 0.00 | 0.04 | 0.12 | 0.07 |

**F. CLASS vs. Teach (N=145)**

| | | Teach | | | | | |
|---|---|---|---|---|---|---|---|
| | | Share of time teaching | Share of time teaching and learning | Classroom culture | Instruction | Socioemotional Skills | Overall Teach |
| **CLASS** | Emotional Support | 0.26 | 0.25 | 0.18 | 0.25 | 0.20 | **0.32** |
| | Classroom Organisation | **0.32** | 0.24 | 0.14 | 0.16 | 0.07 | 0.17 |
| | Instructional Support | 0.24 | 0.24 | 0.14 | **0.31** | 0.23 | **0.37** |
| | Overall CLASS | **0.31** | 0.29 | 0.18 | 0.30 | 0.21 | **0.36** |
| | Student Engagement | **0.31** | 0.25 | 0.17 | 0.30 | 0.23 | **0.37** |

Note: Correlation coefficients above 0.3 are in bold; correlation coefficients above 0.15 are generally statistically significantly different from 0 (at the 5 percent level).

**Table 11: Principal Components Analysis of Level 1 Variables (N=107)**

| | Summary Statistics | | | | Coefficients for first five principal components | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Eigen-value | Diff-erence | Prop. | Variable | Comp1 | Comp2 | Comp3 | Comp4 | Comp5 |
| Comp 1 | **3.86** | 1.97 | 0.23 | **SDI variables** | | | | | |
| Comp 2 | **1.89** | 0.11 | 0.11 | Share of time teaching and learning | 0.22 | **0.39** | 0.15 | 0.11 | -0.28 |
| Comp 3 | **1.78** | 0.17 | 0.10 | Good Teacher Demeanor | 0.16 | -0.03 | 0.23 | **0.42** | **-0.30** |
| Comp 4 | **1.61** | 0.37 | 0.09 | Good Pedagogical Practice | 0.14 | -0.21 | 0.07 | **0.40** | -0.20 |
| Comp 5 | **1.24** | 0.16 | 0.07 | Classroom environment | 0.04 | -0.03 | **0.57** | **-0.39** | -0.12 |
| Comp 6 | **1.09** | 0.09 | 0.06 | Materials and infrastructure | 0.08 | 0.28 | 0.21 | 0.13 | **0.35** |
| Comp 7 | **1.00** | 0.25 | 0.06 | Use of materials | 0.16 | -0.13 | 0.22 | **0.37** | -0.02 |
| Comp 8 | 0.75 | 0.06 | 0.04 | **Stallings variables** | | | | | |
| Comp 9 | 0.69 | 0.06 | 0.04 | Share of time teaching and learning | 0.23 | **0.48** | 0.06 | 0.01 | -0.09 |
| Comp 10 | 0.63 | 0.04 | 0.04 | Materials and infrastructure | 0.14 | 0.22 | 0.02 | 0.05 | **0.55** |
| | | | | Classroom environment | 0.08 | -0.16 | **0.58** | **-0.33** | -0.02 |
| | | | | **CLASS variables** | | | | | |
| | | | | Emotional Support | **0.41** | -0.17 | -0.17 | -0.25 | -0.04 |
| | | | | Classroom Organisation | 0.30 | 0.26 | -0.06 | -0.18 | 0.05 |
| | | | | Instructional Support | **0.41** | -0.19 | -0.21 | -0.15 | -0.10 |
| | | | | Student Engagement | **0.42** | -0.16 | -0.21 | -0.19 | -0.08 |
| | | | | **Teach variables** | | | | | |
| | | | | Share of time teaching and learning | 0.24 | 0.29 | -0.06 | 0.03 | -0.09 |
| | | | | Classroom culture | 0.21 | -0.12 | 0.10 | 0.11 | **0.52** |
| | | | | Instruction | 0.26 | -0.10 | 0.08 | 0.24 | 0.06 |
| | | | | Socioemotional Skills | 0.18 | **-0.37** | 0.10 | 0.10 | 0.21 |

Note: Eigen values greater than 1 and component loadings greater than 0.3 (in absolute value) are in bold.

**Table 12: SDI: Student test scores regressed on classroom observation variables (level 1) and teacher test scores**

| | All teachers | | | | Teachers with test score <=0 | | | | Teachers with test score >0 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| Share of time teaching | 0.9511*** | 0.332 | 0.9780*** | 0.331 | 2.0179*** | 0.869* | 2.1418*** | 1.384*** | 0.4731 | 0.497** | 0.6670* | 0.532 |
| | (0.2988) | (0.273) | (0.2885) | (0.274) | (0.5006) | (0.499) | (0.5465) | (0.488) | (0.3697) | (0.244) | (0.3529) | (0.325) |
| Share of time lost | -0.9479** | -0.293 | -0.9438** | -0.204 | -1.0031* | 0.351 | -1.1645** | -0.224 | -0.8813 | -0.478 | -0.7651 | -0.511 |
| | (0.4329) | (0.422) | (0.3975) | (0.402) | (0.5166) | (0.651) | (0.5056) | (0.664) | (0.8938) | (0.526) | (0.7635) | (0.593) |
| Good Teacher Demeanor | -0.2982 | -0.287 | | | 0.1310 | 1.278** | | | -0.3142 | -0.886** | | |
| | (0.4045) | (0.366) | | | (0.6488) | (0.579) | | | (0.6050) | (0.376) | | |
| Good Pedagogical Practice | -0.6219 | -0.837* | | | -1.0111 | -1.349** | | | 0.0417 | -0.490 | | |
| | (0.4646) | (0.428) | | | (0.6522) | (0.554) | | | (0.8876) | (0.487) | | |
| Classroom Environment | -0.1665 | -0.204 | | | -0.0051 | -0.0585 | | | -0.4918* | -0.326** | | |
| | (0.1875) | (0.166) | | | (0.2012) | (0.149) | | | (0.2754) | (0.158) | | |
| Availability of materials | 1.0799 | 1.223* | | | 0.5545 | 0.206 | | | 1.7949** | 2.279*** | | |
| | (0.7206) | (0.633) | | | (1.0664) | (0.723) | | | (0.8268) | (0.672) | | |
| Use of materials | -0.2564 | 0.223 | | | -0.3344 | -0.315 | | | -0.3834 | 0.426* | | |
| | (0.3118) | (0.291) | | | (0.5303) | (0.497) | | | (0.3993) | (0.232) | | |
| SDI PC1 | | | -0.1311*** | -0.0923** | | | -0.1106 | -0.0284 | | | -0.1242* | -0.120** |
| | | | (0.0437) | (0.0394) | | | (0.0682) | (0.0654) | | | (0.0619) | (0.0455) |
| SDI PC2 | | | 0.0184 | 0.0186 | | | 0.0251 | 0.0130 | | | -0.0110 | 0.0284 |
| | | | (0.0545) | (0.0479) | | | (0.0778) | (0.0634) | | | (0.0565) | (0.0453) |
| Teacher Test Score | -0.0143 | 0.0298 | 0.0012 | 0.0511 | 0.2070** | 0.257*** | 0.1926** | 0.213*** | -0.4860* | -0.310 | -0.3753 | -0.280 |
| | (0.1097) | (0.0719) | (0.1045) | (0.0770) | (0.0959) | (0.0667) | (0.0747) | (0.0587) | (0.2639) | (0.225) | (0.3218) | (0.310) |
| Observations | 4,619 | 4,586 | 4,619 | 4,586 | 2,333 | 2,300 | 2,333 | 2,300 | 2,286 | 2,286 | 2,286 | 2,286 |
| R-squared | 0.057 | 0.245 | 0.049 | 0.229 | 0.104 | 0.303 | 0.100 | 0.288 | 0.083 | 0.322 | 0.050 | 0.270 |
| Controls | NO | YES | NO | YES | NO | YES | NO | YES | NO | YES | NO | YES |
| Teacher Test | YES | YES | YES | YES | YES | YES | YES | YES | YES | YES | YES | YES |
| F-Test: Observation vars. | 0.003 | 0.029 | 0.005 | 0.202 | 0.001 | 0.024 | 0.000 | 0.074 | 0.003 | 0.000 | 0.155 | 0.025 |
| F-Test: Student vars. | | 0.058 | | 0.037 | | 0.000 | | 0.000 | | 0.009 | | 0.002 |
| F-Test: Household vars. | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 |
| F-Test: Teacher vars. | | 0.635 | | 0.614 | | 0.122 | | 0.391 | | 0.001 | | 0.077 |
| F-Test: School vars. | | 0.000 | | 0.000 | | 0.029 | | 0.045 | | 0.000 | | 0.000 |

Note: Standard errors clustered at the teacher level in parentheses. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

**Table 13: Stallings: Student test scores regressed on classroom observation variables (level 1) and teacher test scores**

| | All teachers | | | | Teachers with test score <=0 | | | | Teachers with test score >0 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| Share of time teaching | -0.1878 | -0.425 | -0.1921 | -0.422 | -0.0836 | -0.353 | -0.0661 | -0.393 | -0.2405 | -0.109 | -0.2995 | -0.150 |
| | (0.3188) | (0.267) | (0.3202) | (0.267) | (0.6409) | (0.643) | (0.6465) | (0.633) | (0.5556) | (0.334) | (0.5335) | (0.334) |
| Share of time lost | -0.7645 | -0.640 | -0.8052 | -0.633 | -1.0848 | -1.778*** | -1.1037 | -1.788*** | -0.3623 | -0.653 | -0.4297 | -0.642 |
| | (0.5050) | (0.460) | (0.5102) | (0.458) | (0.9959) | (0.634) | (1.0084) | (0.654) | (0.6527) | (0.584) | (0.6837) | (0.599) |
| Availability of materials | 0.4661 | -0.178 | | | -0.0566 | -1.143 | | | 1.5375 | 0.978 | | |
| | (0.8967) | (0.705) | | | (1.0873) | (0.761) | | | (1.1090) | (0.841) | | |
| Classroom environment | -0.1510 | 0.0446 | | | -0.1722 | 0.0507 | | | -0.1290 | -0.0661 | | |
| | (0.2204) | (0.208) | | | (0.3036) | (0.238) | | | (0.2762) | (0.185) | | |
| Stallings PC1 | | | -0.0074 | -0.000460 | | | -0.0397 | -0.0507 | | | 0.0566 | 0.0393 |
| | | | (0.0709) | (0.0544) | | | (0.0857) | (0.0597) | | | (0.0812) | (0.0562) |
| Teacher Test Score | -0.0368 | 0.0171 | -0.0487 | 0.0201 | 0.1397 | 0.00641 | 0.1288 | 0.0232 | -0.3774 | -0.343 | -0.4326 | -0.369 |
| | (0.1106) | (0.0767) | (0.1101) | (0.0775) | (0.0926) | (0.0751) | (0.0868) | (0.0789) | (0.3089) | (0.276) | (0.3204) | (0.292) |
| Observations | 4,619 | 4,586 | 4,619 | 4,586 | 2,333 | 2,300 | 2,333 | 2,300 | 2,286 | 2,286 | 2,286 | 2,286 |
| R-squared | 0.027 | 0.230 | 0.024 | 0.229 | 0.050 | 0.295 | 0.049 | 0.293 | 0.040 | 0.259 | 0.030 | 0.257 |
| Controls | NO | YES | NO | YES | NO | YES | NO | YES | NO | YES | NO | YES |
| Teacher Test | YES | YES | YES | YES | YES | YES | YES | YES | YES | YES | YES | YES |
| F-Test: Observation vars. | 0.434 | 0.381 | 0.421 | 0.254 | 0.691 | 0.049 | 0.625 | 0.044 | 0.584 | 0.454 | 0.737 | 0.418 |
| F-Test: Student vars. | | 0.011 | | 0.011 | | 0.000 | | 0.000 | | 0.004 | | 0.004 |
| F-Test: Household vars. | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 |
| F-Test: Teacher vars. | | 0.116 | | 0.113 | | 0.016 | | 0.018 | | 0.010 | | 0.005 |
| F-Test: School vars. | | 0.000 | | 0.001 | | 0.062 | | 0.114 | | 0.000 | | 0.000 |

Note: Standard errors clustered at the teacher level in parentheses. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

**Table 14: CLASS: Student test scores regressed on classroom observation variables (level 1) and teacher test scores**

| | All teachers | | | | Teachers with test score <=0 | | | | Teachers with test score >0 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| Emotional Support | -0.0809 | 0.0186 | | | -0.3543 | -0.105 | | | 0.0109 | -0.181 | | |
| | (0.1742) | (0.126) | | | (0.2658) | (0.196) | | | (0.1869) | (0.143) | | |
| Classroom Organisation | 0.1189 | 0.122 | | | 0.3015 | 0.142 | | | -0.0252 | 0.205* | | |
| | (0.1308) | (0.113) | | | (0.2390) | (0.198) | | | (0.1297) | (0.117) | | |
| Instructional Support | 0.2329 | 0.149 | | | 0.2971 | 0.153 | | | 0.2603 | 0.263* | | |
| | (0.2008) | (0.145) | | | (0.3039) | (0.211) | | | (0.2433) | (0.152) | | |
| Student Engagement | -0.0376 | -0.0239 | | | -0.0236 | 0.120 | | | 0.0096 | -0.131 | | |
| | (0.1234) | (0.0844) | | | (0.2049) | (0.176) | | | (0.1472) | (0.118) | | |
| CLASS PC1 | | | 0.0736 | 0.0845*** | | | -0.0123 | 0.103* | | | 0.1352** | 0.0336 |
| | | | (0.0470) | (0.0294) | | | (0.0735) | (0.0539) | | | (0.0513) | (0.0499) |
| Teacher Test Score | -0.0567 | -0.00814 | -0.0444 | -0.00213 | 0.2415*** | 0.143** | 0.2218*** | 0.119* | -0.3509 | -0.372 | -0.3482 | -0.359 |
| | (0.1178) | (0.0796) | (0.1075) | (0.0784) | (0.0681) | (0.0689) | (0.0687) | (0.0691) | (0.2811) | (0.315) | (0.2737) | (0.297) |
| Observations | 4,619 | 4,586 | 4,619 | 4,586 | 2,333 | 2,300 | 2,333 | 2,300 | 2,286 | 2,286 | 2,286 | 2,286 |
| R-squared | 0.029 | 0.235 | 0.023 | 0.234 | 0.054 | 0.290 | 0.034 | 0.288 | 0.069 | 0.265 | 0.060 | 0.252 |
| Controls | NO | YES | NO | YES | NO | YES | NO | YES | NO | YES | NO | YES |
| Teacher Test | YES | YES | YES | YES | YES | YES | YES | YES | YES | YES | YES | YES |
| F-Test: Observation vars. | 0.441 | 0.065 | 0.121 | 0.005 | 0.507 | 0.266 | 0.868 | 0.063 | 0.089 | 0.058 | 0.011 | 0.504 |
| F-Test: Student vars. | | 0.020 | | 0.024 | | 0.000 | | 0.000 | | 0.001 | | 0.000 |
| F-Test: Household vars. | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 |
| F-Test: Teacher vars. | | 0.161 | | 0.167 | | 0.181 | | 0.169 | | 0.048 | | 0.144 |
| F-Test: School vars. | | 0.000 | | 0.000 | | 0.001 | | 0.002 | | 0.000 | | 0.000 |

Note: Standard errors clustered at the teacher level in parentheses. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

**Table 15: Teach: Student test scores regressed on classroom observation variables (level 1) and teacher test scores**

| | All teachers | | | | Teachers with test score <=0 | | | | Teachers with test score >0 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| Share of time teaching | 0.7390** | 0.343 | 0.5026 | 0.137 | 1.7094*** | 1.226*** | 1.1430** | 0.656 | -0.3998 | -0.307 | -0.4247 | -0.285 |
| | (0.3668) | (0.343) | (0.3759) | (0.321) | (0.3174) | (0.366) | (0.4249) | (0.416) | (0.4583) | (0.516) | (0.4532) | (0.490) |
| Share of time lost | -0.4146 | -0.486 | -0.5124 | -0.485 | 0.8190 | 1.140 | -0.3081 | -0.0645 | -0.3392 | -0.475 | -0.3212 | -0.451 |
| | (0.4780) | (0.460) | (0.3647) | (0.391) | (1.8768) | (1.574) | (1.9505) | (1.600) | (0.3451) | (0.312) | (0.3587) | (0.366) |
| Classroom culture | 0.0837 | 0.0407 | | | -0.0383 | -0.407 | | | 0.2768 | 0.263 | | |
| | (0.1924) | (0.184) | | | (0.2901) | (0.268) | | | (0.2644) | (0.209) | | |
| Instruction | -0.3458 | -0.0894 | | | -0.9166*** | -0.498** | | | 0.3884 | 0.472* | | |
| | (0.2135) | (0.198) | | | (0.1912) | (0.192) | | | (0.3346) | (0.250) | | |
| Socioemotional Skills | 0.6695*** | 0.483*** | | | 1.1166*** | 1.239*** | | | 0.4068* | 0.134 | | |
| | (0.1911) | (0.179) | | | (0.2669) | (0.223) | | | (0.2197) | (0.170) | | |
| Teach PC1 | | | 0.0880* | 0.0795* | | | -0.0064 | 0.0353 | | | 0.1810*** | 0.141** |
| | | | (0.0497) | (0.0416) | | | (0.0798) | (0.0694) | | | (0.0664) | (0.0586) |
| Teacher Test Score | -0.0073 | 0.0189 | -0.0198 | 0.0275 | 0.1585** | 0.138*** | 0.2264*** | 0.177*** | -0.1810 | -0.328 | -0.1812 | -0.249 |
| | (0.0913) | (0.0777) | (0.1055) | (0.0782) | (0.0661) | (0.0478) | (0.0741) | (0.0620) | (0.3231) | (0.323) | (0.3058) | (0.320) |
| Observations | 4,619 | 4,586 | 4,619 | 4,586 | 2,333 | 2,300 | 2,333 | 2,300 | 2,286 | 2,286 | 2,286 | 2,286 |
| R-squared | 0.0618 | 0.237 | 0.0307 | 0.229 | 0.1699 | 0.326 | 0.0691 | 0.282 | 0.0678 | 0.275 | 0.0655 | 0.270 |
| Controls | NO | YES | NO | YES | NO | YES | NO | YES | NO | YES | NO | YES |
| Teacher Test | YES | YES | YES | YES | YES | YES | YES | YES | YES | YES | YES | YES |
| F-Test: Observation vars. | 0.0173 | 0.138 | 0.101 | 0.111 | 0.000 | 0.000 | 0.0758 | 0.425 | 0.0566 | 0.0159 | 0.0176 | 0.00330 |
| F-Test: Student vars. | | 0.0382 | | 0.0479 | | 0.000 | | 0.000 | | 0.000594 | | 0.000492 |
| F-Test: Household vars. | | 0 | | 0 | | 0.000 | | 0.000 | | 0 | | 0 |
| F-Test: Teacher vars. | | 0.179 | | 0.195 | | 0.132 | | 0.463 | | 0.0224 | | 0.0108 |
| F-Test: School vars. | | 0.00310 | | 0.000956 | | 0.00374 | | 0.00577 | | 0.000148 | | 0.000 |

Note: Standard errors clustered at the teacher level in parentheses. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

**Table 16: All instruments combined: Student test scores regressed on classroom observation variables (level 1) and teacher test scores**

| | All teachers | | | | Teachers with test score <=0 | | | | Teachers with test score >0 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| Share of time teaching (SDI) | 0.9962*** | 0.311 | 1.2284*** | 0.447* | 1.7596*** | 2.064*** | 2.5704*** | 2.299*** | 0.5681* | 0.542** | 0.4511 | 0.492* |
| | (0.3102) | (0.258) | (0.2724) | (0.241) | (0.5458) | (0.742) | (0.5214) | (0.444) | (0.3185) | (0.254) | (0.3090) | (0.292) |
| Share of time lost (SDI) | -0.9036** | -0.114 | -0.9973*** | -0.0736 | -0.8839** | -0.312 | -1.3845*** | -0.516 | -0.6778 | -0.265 | -0.6568 | 0.182 |
| | (0.3548) | (0.379) | (0.3636) | (0.369) | (0.3524) | (0.684) | (0.3865) | (0.533) | (0.5783) | (0.627) | (0.6303) | (0.624) |
| Good Teacher Demeanor (SDI) | -0.1249 | -0.139 | | | 0.2360 | 0.961 | | | -0.2602 | -0.969** | | |
| | (0.3847) | (0.327) | | | (0.5541) | (0.624) | | | (0.4901) | (0.400) | | |
| Good Pedagogical Practise (SDI) | -0.6965 | -1.071** | | | -1.5701** | -1.599*** | | | -0.2187 | -0.483 | | |
| | (0.4906) | (0.420) | | | (0.6165) | (0.379) | | | (0.7559) | (0.412) | | |
| Classroom Environment (SDI) | -0.2646 | -0.298** | | | -0.0817 | -0.153 | | | -0.6463** | -0.352** | | |
| | (0.1652) | (0.138) | | | (0.1777) | (0.118) | | | (0.2409) | (0.148) | | |
| Availability of materials (SDI) | 1.2443* | 1.122** | | | 1.4295 | 0.171 | | | 2.0536* | 2.098** | | |
| | (0.7018) | (0.528) | | | (0.9693) | (0.623) | | | (1.0212) | (0.808) | | |
| Use of materials (SDI) | -0.2610 | 0.351 | | | -0.4364 | -0.110 | | | -0.3071 | 0.333 | | |
| | (0.3094) | (0.270) | | | (0.3149) | (0.377) | | | (0.3560) | (0.286) | | |
| Emotional Support (CLASS) | -0.0681 | -0.0133 | | | -0.2340 | 0.209 | | | 0.2271 | 0.0991 | | |
| | (0.1471) | (0.126) | | | (0.2386) | (0.235) | | | (0.2172) | (0.158) | | |
| Classroom Organisation (CLASS | -0.0107 | 0.0871 | | | 0.0926 | -0.348* | | | -0.2296* | 0.0476 | | |
| | (0.1342) | (0.117) | | | (0.1950) | (0.203) | | | (0.1342) | (0.115) | | |
| Instructional Support (CLASS) | 0.1843 | 0.159 | | | 0.4169* | -0.500** | | | 0.0030 | -0.0342 | | |
| | (0.1744) | (0.140) | | | (0.2232) | (0.244) | | | (0.2352) | (0.140) | | |
| Student Engagement (CLASS) | -0.0098 | 0.0381 | | | -0.1626 | 0.598*** | | | 0.0330 | -0.0377 | | |
| | (0.1244) | (0.0862) | | | (0.1784) | (0.163) | | | (0.1269) | (0.0866) | | |
| Classroom culture (Teach) | 0.0324 | -0.0248 | | | 0.3869 | -0.758*** | | | 0.1369 | 0.233 | | |
| | (0.2220) | (0.170) | | | (0.2610) | (0.253) | | | (0.1869) | (0.152) | | |
| Instruction (Teach) | -0.4386** | -0.163 | | | -0.8424*** | -0.146 | | | 0.4375* | 0.454*** | | |
| | (0.2121) | (0.178) | | | (0.2526) | (0.248) | | | (0.2276) | (0.163) | | |
| Socioemotional Skills (Teach) | 0.6576*** | 0.288* | | | 0.8604*** | 1.180*** | | | 0.3802** | 0.0251 | | |
| | (0.1912) | (0.163) | | | (0.2919) | (0.258) | | | (0.1825) | (0.158) | | |

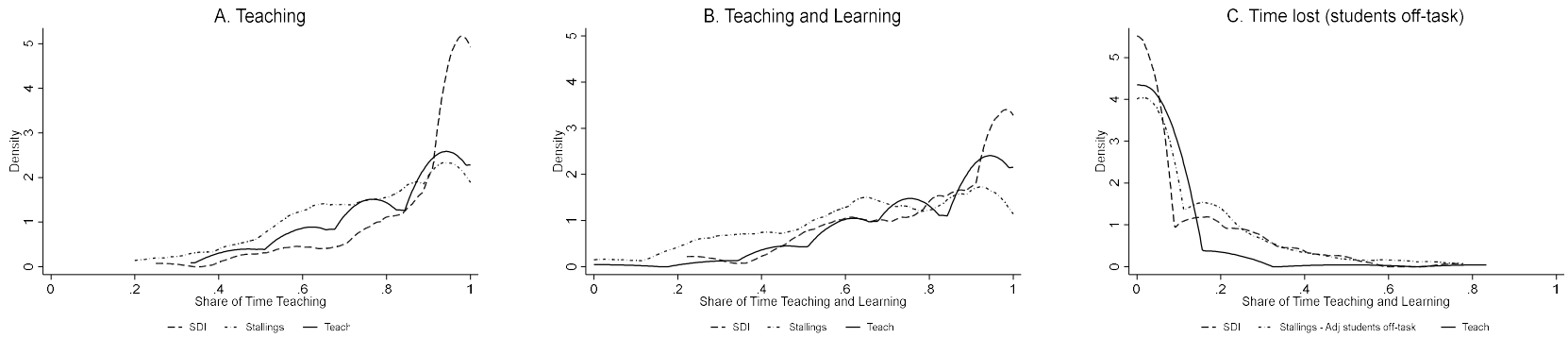
continued…

**Table 16 continued: All instruments combined: Student test scores regressed on classroom observation variables (level 1) and teacher test scores**

| | All teachers | | | | Teachers with test score <=0 | | | | Teachers with test score >0 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| PC1: "general good teaching practice" | | | -0.0097 | 0.0277 | | | -0.0610 | 0.0693 | | | 0.0797 | 0.00590 |
| | | | (0.0375) | (0.0293) | | | (0.0428) | (0.0457) | | | (0.0627) | (0.0561) |
| PC2: "good SDI/bad CLASS" | | | -0.1616*** | -0.118*** | | | -0.1474** | -0.131** | | | -0.1369** | -0.117*** |
| | | | (0.0395) | (0.0350) | | | (0.0702) | (0.0585) | | | (0.0538) | (0.0388) |
| PC3: "good classroom atmosphere" | | | 0.0496 | 0.0347 | | | 0.1092 | -0.0412 | | | 0.0123 | 0.158*** |
| | | | (0.0491) | (0.0409) | | | (0.0962) | (0.0570) | | | (0.0325) | (0.0411) |
| PC4: "poor support to socio-emotional skills development" | | | -0.1496*** | -0.0692 | | | -0.1961** | -0.212*** | | | -0.1050* | -0.0452 |
| | | | (0.0557) | (0.0496) | | | (0.0775) | (0.0740) | | | (0.0610) | (0.0541) |
| Teacher Test Score | -0.0281 | -0.0211 | 0.0075 | 0.0226 | 0.1541 | 0.196*** | 0.1737** | 0.140** | -0.2053 | -0.315 | -0.0967 | -0.216 |
| | (0.0876) | (0.0637) | (0.0843) | (0.0701) | (0.0921) | (0.0636) | (0.0768) | (0.0528) | (0.2465) | (0.254) | (0.3156) | (0.300) |
| Observations | 4,619 | 4,586 | 4,619 | 4,586 | 2,333 | 2,300 | 2,333 | 2,300 | 2,286 | 2,286 | 2,286 | 2,286 |
| R-squared | 0.109 | 0.272 | 0.082 | 0.247 | 0.191 | 0.367 | 0.134 | 0.324 | 0.151 | 0.337 | 0.098 | 0.285 |
| Controls | NO | YES | NO | YES | NO | YES | NO | YES | NO | YES | NO | YES |
| Teacher Test | YES | YES | YES | YES | YES | YES | YES | YES | YES | YES | YES | YES |
| F-Test: Observation vars. | 0.000 | 0.000 | 0.000 | 0.008 | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.004 | 0.000 |
| F-Test: Student vars. | | 0.010 | | 0.022 | | 0.000 | | 0.000 | | 0.003 | | 0.004 |
| F-Test: Household vars. | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 |
| F-Test: Teacher vars. | | 0.419 | | 0.521 | | 0.002 | | 0.026 | | 0.007 | | 0.002 |
| F-Test: School vars. | | 0.000 | | 0.001 | | 0.000 | | 0.004 | | 0.000 | | 0.000 |

Note: Standard errors clustered at the teacher level in parentheses. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

**Appendix 1 Figure 1: Distribution of share of time variables (Common observations only)**



**Appendix 1 Table 1: SDI Classroom Observation Instrument elements**

| **Time spent on various activities** | | |
|---|---|---|
| Teacher in class - teaching<br>Teacher interacts with all children as a group<br>Teacher interacts with a small group of children<br>Teacher interacts with children one-on-one<br>Teacher reads or lectures to the pupils (pupils only listen)<br>Teacher supervises pupil(s) writing on the board<br>Teacher leads kinesthetic group learning activity<br>Teacher writing on blackboard<br>Teacher listening to pupils recite/read<br>Teacher waiting for pupils to complete task<br>Teacher testing students in class | | Teacher - teaching |
| Teacher maintaining discipline in class<br>Teacher doing paperwork<br>Teacher in class - not teaching<br>Teacher not in class - learning activity ongoing<br>Teacher not in class - no learning activity ongoing<br>Break | | Teacher - not teaching |
| Number of pupils off task (every 5 minutes) | | Used to adjust teaching time |
| **Good teacher demeanor** | Score | |
| Teacher demeanor<br>Was the teacher either sitting or standing in front of the class at any time?<br>Did the teacher visit individual children?<br>* How many pupils did the teacher go to individually?<br>Did the teacher call pupils by name while teaching?<br>* How many pupils did the teacher call by name?<br>Was the teacher smiling, laughing, or joking with pupils?<br>Did the teacher hit, pinch, or slap a pupil? | <br>1=mostly sitting<br>1=yes<br>1= >20%<br>1=yes<br>1= >15%<br>1=yes<br>1=no | Good teacher demeanor = average |

continued…

**Appendix 1 Table 1 continued: SDI Classroom Observation Instrument elements**

| | | |
|---|---|---|
| **Good pedagogical practices** | | |
| Teacher asking questions | | Good pedagogical practices = average |
| Did the teacher ask questions that required learners to recall information? | 1=yes | |
| Did the teacher ask learners to carry out a task which allowed them to demo | 1=yes | |
| Did the teacher ask questions that required learners to apply information to | 1=yes | |
| Did the teacher ask questions which required learners to use their creativity | 1=yes | |
| Feedback | | |
| Did the teacher give feedback or praise, moral strengthening, and/or encour | .33=never, .66=once, 1 = more than once | |
| Did the teacher give feedback that was correcting a mistake? | .33=never, .66=once, 1 = more than once | |
| Did the teacher give feedback that was scolding at a mistake? | .33=more than once, .66=once, 1 = never | |
| Introducing and summarizing lesson | | |
| Did the teacher introduce the lesson at the start of the class? | 1=yes | |
| Did the teacher summarize the lesson at the end of the class? | 1=yes | |
| Homework | | |
| Did the teacher assign homework to the class? | 1=yes | |
| Did the teacher review or collect homework from the class? | 1=yes | |
| Language | | |
| Did the teacher use the local language as a medium of instruction? (languag | 1=yes | |
| Local Information | | |
| Did the teacher use local information from the community to make learning | 1=yes | |
| **Classroom environment** | | |
| Was pupils' work displayed on the walls? | 1=yes | Classroom environment = average |
| Other than pupils' work, were there other materials, such as, charts displayed on | 1=yes | |
| | | |
| | | |
| **Availability of materials and classroom infrastructure** | | |
| Is there a "corner library" in the class or additional available books for pupils? | 1=yes | Availability of materials and classroom infrastructure = average |
| Is there a blackboard and/or whiteboard in the class? | 1=yes | |
| Is there chalk or marker to write on the board available during the lesson? | 1=yes | |
| How many pupils were not sitting at desks? | 1 = <3% | |
| Does the classroom have a working electricity connection (e.g., electric light)? | 1=yes | |
| How would you classify the hygiene in the classroom? | 1 = reasonably or extremely clean | |
| Does the blackboard have sufficient contrast for reading what is written on the b | 1=yes | |
| **Use of materials** | | |
| Was the text book used by the teacher? | 1=yes | Use of materials = average |
| Did the teacher write on the black board? | 1=yes | |
| Did any pupils write on the black board? | 1=yes | |

**Appendix 1 Table 2: Stallings Observation System**

| Lesson Time Use | Level 1 variables |
|---|---|
| Active Instruction | Teacher teaching |
| Reading | |
| Instruct Explain | |
| Discussion | |
| Practice Drill | |
| Passive Instruction | |
| Monitoring Seatwork | |
| Monitoring Copying | |
| Organizing Management | |
| Giving Assignements | Teacher not teaching |
| Managing With Students | |
| Disciplining Students | |
| Managing Alone | |
| Teacher off task | |
| Student off task rate | Used to adjust teaching time |
| Being Disciplined | |
| Socializing | |
| Uninvolved | |
| Class characteristics | Not used |
| Number of students | |
| Availability of materials and classroom infrastructure | Availability of materials |
| Blackboard in the classroom | |
| Textbook/other printed material | |
| Notebook/Writing material | |
| Classroom environment | Classroom environment |
| Display of charts, pictures, maps on the wall | |
| School Uniform | |

**Appendix 1 Table 3: Classroom Assessment Scoring System**

| Dimension (Level 2) | Indicators | Domain (Level 1) |
|---|---|---|
| Positive Climate | | Emotional Support |
| | Relationships | |
| | Positive Affect | |
| | Positive Communications | |
| | Respect | |
| Teacher Sensitivity | | |
| | Awareness | |
| | Responsiveness to Academic and Social/Emotional needs and cues | |
| | Effectiveness In Addressing Problems | |
| | Student Comfort | |
| Regard for Adolescent Perspective | | |
| | Flexibility And Student Focus | |
| | Connections To Current Life | |
| | Support For Autonomy And Leadership | |
| | Meaningful Peer Interactions | |
| Negative Climate | | Classroom Organization |
| | Negative Affect | |
| | Punitive Control | |
| | Disrespect | |
| Behavior Management | | |
| | Clear Expectations | |
| | Proactive | |
| | Effective Redirection Of Misbehavior | |
| | Student Behavior | |
| Productivity | | |
| | Maximizing Learning Time | |
| | Routines | |
| | Transitions | |
| | Preparation | |
| Instructional Learning Format | | Instructional support |
| | Learning Targets/Organization | |
| | Variety Of Modalities, Strategies, and materials | |
| | Active Facilitation | |
| | Effective Engagement | |
| Content Understanding | | |
| | Depth Of Understanding | |
| | Communication Of Concepts And Procedures | |
| | Background Knowledge And Misconceptions | |
| | Transmission Of Content Knowledge And Procedures | |
| | Opportunity For Practice Of Procedures And Skills | |
| Analysis and Inquiry | | |
| | Facilitation Of Higher-Order Thinking | |
| | Opportunities For Novel Application | |
| | Metacognition | |
| Quality of Feedback | | |
| | Feedback Loops | |
| | Scaffolding | |
| | Building On Student Responses | |
| | Encouragement And Affirmation | |
| Instuctional Dialogue | | |
| | Cumulative Content-Driven Exchanges | |
| | Distributed Talk | |
| | Facilitation Strategies | |
| | Active Engagement | Student Engagement |

**Appendix 1 Table 4: Teach Classroom Observation Tool**

| Teacher Provides Learning Activity To Most Students (Yes/no) | Teacher teaching |
|---|---|
| Students Are On Task (Low/Medium/High) | Used to adjust teaching time |
| Dimension (Level 2) | Area (Level 1) |
| Supportive Learning Environment | |
| The Teacher Treats All Students Respectfully | |
| The Teacher Uses Positive Language With Students | |
| The Teacher Resonds To Students' Needs | |
| The Teacher Does Not Exhibit Gender Bias And Challenges Gender Sterotypes In The Classroom | Classroom Culture |
| Positive Behaviorial Expectations | |
| The Teacher Sets Clear Behavioral Expectations For Classroom Activities | |
| The Teacher Acknowledges Positive Student Behavior | |
| The Teacher Redirects Misbehavior And Focuses On The Expected Behavior, Rather Than The Undesired Behavior | |
| Lesson Facilitation | |
| The Teacher Explicitly Articulates The Objectives Of The Lesson And Relates Classroom Activities To The Objectives | |
| The Teacher'S Explanation Of Content Is Clear | |
| The Teacher Makes Connections In The Lesson That Relate To Other Content Knowledge Or Students' Daily Lives | |
| The Teacher Models By Enacting Or Thinking Aloud | |
| Checks For Understanding | |
| The Teacher Uses Questions, Prompts Or Other Strategies To Determine Students' Level Of Understanding | |
| The Teacher Monitors Most Students During Independent/Group Work | Instruction |
| The Teacher Adjusts Teaching To The Level Of Students | |
| Feedback | |
| The Teacher Provides Specific Comments Or Prompts That Help Clarify Students' Misunderstandings | |
| The Teacher Provides Specific Comments Or Prompts That Help Identify Students' Successes | |
| Critical Thinking | |
| The Teacher Asks Open-Ended Questions | |
| The Teacher Provides Thinking Tasks | |
| The Students Ask Open-Ended Questions Or Perform Thinking Tasks | |
| Autonomy | |
| The Teacher Provides Students With Choices | |
| The Teacher Provides Students With Opportunities To Take On Roles In The Classroom | |
| The Students Volunteer To Participate In The Classroom | |
| Perseverance | |
| The Teacher Acknowledges Students' Efforts | Socioemotional Skills |
| The Teacher Has A Positive Attitude Towards Students' Challenges | |
| The Teacher Encourages Goal Setting | |
| Social & Collaborative Skills | |
| The Teacher Promotes Students' Collaboration Through Peer Interaction | |
| The Teacher Promotes Students' Interpersonal Skills | |
| Students Collaborate With One Another Through Peer Interaction | |

**Appendix 1 Table 5: Summary statistics on classroom conditions during SDI observations of the**

| | All (N=400) | | Rural (N=271) | | Urban (N=129) | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD |
| Blackboard in class (1=yes) | 0.98 | (0.13) | 0.99 | (0.10) | 0.97 | (0.17) |
| Blackboard has sufficient contrast for reading (1=yes) | 0.90 | (0.30) | 0.87 | (0.34) | 0.97 | (0.17) |
| Chalk available during the lesson (1=yes) | 0.97 | (0.18) | 0.97 | (0.18) | 0.97 | (0.17) |
| Classroom has a working electricity connection (1=yes) | 0.04 | (0.20) | 0.03 | (0.17) | 0.07 | (0.26) |
| Hygiene in class (1 = reasonably or extremely clean) | 0.70 | (0.46) | 0.65 | (0.48) | 0.80 | (0.40) |
| Pupils work displayed on the the wall | 0.20 | (0.40) | 0.21 | (0.41) | 0.19 | (0.39) |
| Charts displayed on the walls | 0.28 | (0.45) | 0.27 | (0.44) | 0.31 | (0.46) |
| Corner Library | 0.03 | (0.18) | 0.03 | (0.18) | 0.03 | (0.17) |

Source: SDI 2016

**Appendix 1 Table 6: Summary statistics on selected teacher practices observed during classroom observations in the nationally representative sample of 400 schools**

| | All (N=400) | | Rural (N=271) | | Urban (N=129) | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD |
| **Use of materials** | | | | | | |
| Textbook was used by teacher | 0.87 | (0.33) | 0.88 | (0.32) | 0.85 | (0.36) |
| Teacher writes on the blackboard | 0.99 | (0.10) | 0.99 | (0.10) | 0.99 | (0.09) |
| Pupils write on the blackboard | 0.40 | (0.49) | 0.41 | (0.49) | 0.38 | (0.49) |
| Teacher uses local information to make learning relevant | 0.46 | (0.50) | 0.45 | (0.50) | 0.47 | (0.50) |
| **Teacher demeanor** | | | | | | |
| Teacher mostly standing (as opposed to sitting) | 0.95 | (0.22) | 0.97 | (0.18) | 0.91 | (0.28) |
| Teacher visits children individually | 0.60 | (0.49) | 0.59 | (0.49) | 0.62 | (0.49) |
| Teacher calls pupil by name while teaching | 0.80 | (0.40) | 0.81 | (0.39) | 0.78 | (0.41) |
| Teacher smiling, laughing, or joking with pupils | 0.56 | (0.50) | 0.54 | (0.50) | 0.59 | (0.49) |
| Teacher hit pupil (1=no) | 0.94 | (0.24) | 0.93 | (0.25) | 0.95 | (0.23) |
| **Pedagogical practices** | | | | | | |
| Teacher ask questions that required learners to recall information | 0.73 | (0.44) | 0.73 | (0.45) | 0.74 | (0.44) |
| Teacher ask learners to to demonstrate their understanding | 0.93 | (0.25) | 0.94 | (0.24) | 0.92 | (0.27) |
| Teacher ask questions that required learners to apply information | 0.82 | (0.38) | 0.84 | (0.37) | 0.79 | (0.41) |
| Teacher ask questions requiring learners to use their creativity | 0.75 | (0.43) | 0.73 | (0.44) | 0.79 | (0.41) |
| Teacher gives feedback and/or encouragement to students (scale, 1= | 0.71 | (0.39) | 0.70 | (0.40) | 0.74 | (0.38) |
| Teacher gives feedback that was correcting a mistake (scale, 1=yes) | 0.74 | (0.39) | 0.75 | (0.37) | 0.72 | (0.42) |
| Teachers gives feedback that was scolding a mistake (scale, 1=no) | 0.73 | (0.40) | 0.74 | (0.39) | 0.71 | (0.42) |
| Teacher introduces the lesson at the start of class | 0.90 | (0.31) | 0.88 | (0.33) | 0.93 | (0.26) |
| Teacher summarize the lesson at the end of class | 0.45 | (0.50) | 0.46 | (0.50) | 0.42 | (0.50) |
| Teacher assigns homework | 0.43 | (0.50) | 0.48 | (0.50) | 0.33 | (0.47) |
| Teacher reviews homeworks | 0.24 | (0.43) | 0.27 | (0.44) | 0.18 | (0.38) |
| Teacher uses local langague for instructions | 0.11 | (0.31) | 0.08 | (0.27) | 0.16 | (0.37) |

Source: SDI 2016 classroom observations. Notes: All answers are binary 1=yes 0=no, unless otherwise indicated as scale variables (where 3 categories were rescaled to be between 0 and 1)

**Appendix 1 Table 7: School-level Principal Components Analysis of Level 1 Variables**

| Summary Statistics | | | | Coefficients for first three principal components | | | |
|---|---|---|---|---|---|---|---|
| **SDI (N=104)** | | | | | | | |
| Comp1 | **1.95** | 0.82 | 0.39 | Share of time teaching and learning | - | - | - |
| Comp2 | **1.13** | 0.22 | 0.23 | Good Teacher Demeanor | **0.61** | 0.06 | -0.19 |
| Comp3 | 0.91 | 0.29 | 0.18 | Good Pedagogical Practise | **0.58** | 0.01 | 0.18 |
| | | | | Classroom environment | -0.01 | **0.72** | -0.66 |
| | | | | Materials and infrastructure | -0.01 | **0.69** | **0.70** |
| | | | | Use of materials | **0.54** | -0.06 | 0.02 |
| **Stallings (N=105)** | | | | | | | |
| Comp1 | **1.18** | 0.36 | 0.59 | Share of time teaching and learning | - | - | |
| Comp2 | 0.82 | . | 0.41 | Materials and infrastructure | **0.707** | 0.707 | |
| | | | | Classroom environment | **0.707** | -0.707 | |
| **CLASS (N=96)** | | | | | | | |
| Comp1 | **2.97** | 2.24 | 0.74 | Emotional Support | **0.54** | -0.24 | **0.47** |
| Comp2 | 0.73 | 0.55 | 0.18 | Classroom Organisation | **0.35** | **0.93** | 0.11 |
| Comp3 | 0.18 | 0.05 | 0.04 | Instructional Support | **0.54** | -0.27 | **0.28** |
| | | | | Student Engagement | **0.54** | -0.10 | -0.83 |
| **Teach (N=97)** | | | | | | | |
| Comp1 | **1.47** | 0.61 | 0.49 | Share of time teaching and learning | - | - | - |
| Comp2 | 0.86 | 0.19 | 0.29 | Classroom culture | **0.60** | -0.46 | **0.65** |
| Comp3 | 0.67 | . | 0.22 | Instruction | **0.49** | **0.86** | 0.17 |
| | | | | Socioemotional Skills | **0.63** | -0.22 | -0.74 |

Note: In this analysis, share of time teaching and learning is excluded from the principal components analysis. Eigen values greater than 1 and component loadings greater than 0.3 (in absolute value) are in bold.

**Appendix 1 Table 8: School-level Principal Components Analysis of Level 1 Variables**

| Summary Statistics | | | | Coefficients for first five principal components | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Eigen-value | Diff-erence | Prop. | Variable | Comp1 | Comp2 | Comp3 | Comp4 | Comp5 |
| Comp1 | **3.90** | 2.18 | 0.32 | **SDI variables** | | | | | |
| Comp2 | **1.71** | 0.45 | 0.14 | Share of time teaching and learning | - | - | - | - | - |
| Comp3 | **1.26** | 0.22 | 0.11 | Good Teacher Demeanor | 0.22 | **0.52** | -0.05 | 0.11 | 0.29 |
| Comp4 | **1.05** | 0.06 | 0.09 | Good Pedagogical Practice | 0.25 | **0.49** | -0.09 | 0.14 | -0.08 |
| Comp5 | 0.99 | 0.19 | 0.08 | Classroom environment | 0.04 | -0.08 | **0.43** | 0.22 | **0.82** |
| Comp6 | 0.80 | 0.16 | 0.07 | Materials and infrastructure | 0.06 | -0.11 | **0.64** | **0.43** | **-0.33** |
| Comp7 | 0.64 | 0.08 | 0.05 | Use of materials | 0.21 | **0.51** | 0.09 | -0.11 | -0.05 |
| Comp8 | 0.56 | 0.15 | 0.05 | **Stallings variables** | | | | | |
| Comp9 | 0.42 | 0.04 | 0.03 | Share of time teaching and learning | - | - | - | - | - |
| Comp10 | 0.38 | 0.21 | 0.03 | Materials and infrastructure | - | - | - | - | - |
| | | | | Classroom environment | - | - | - | - | - |
| | | | | **CLASS variables** | | | | | |
| | | | | Emotional Support | **0.42** | -0.28 | -0.19 | 0.06 | 0.10 |
| | | | | Classroom Organisation | **0.31** | -0.05 | 0.18 | 0.20 | **-0.29** |
| | | | | Instructional Support | **0.43** | -0.24 | -0.20 | 0.04 | 0.08 |
| | | | | Student Engagement | **0.43** | -0.24 | -0.19 | 0.08 | 0.04 |
| | | | | **Teach variables** | | | | | |
| | | | | Share of time teaching and learning | - | - | - | - | - |
| | | | | Classroom culture | 0.22 | 0.05 | **0.44** | **-0.41** | -0.10 |
| | | | | Instruction | **0.32** | 0.03 | 0.01 | 0.11 | -0.14 |
| | | | | Socioemotional Skills | 0.22 | -0.10 | 0.23 | **-0.70** | 0.06 |

Note: In this analysis, share of time teaching and learning and redundant variables are excluded from the principal components analysis. Eigen values greater than 1 and component loadings greater than 0.3 (in absolute value) are in bold.

**Appendix 1 Table 9: Student test scores regressed on classroom observation variables (level 1), variable-by-variable**

| | All teachers | | Teachers with test score <=0 | | Teachers with test score >0 | |
|---|---|---|---|---|---|---|
| **SDI** | | | | | | |
| Share of time teaching | 0.408 | -0.092 | 1.610 *** | 1.126 ** | 0.044 | -0.126 |
| | (0.295) | (0.263) | (0.538) | (0.449) | (0.322) | (0.278) |
| Share of time lost | -0.777 * | -0.258 | -0.819 | 0.025 | -0.849 | -0.994 * |
| | (0.417) | (0.431) | (0.513) | (0.598) | (0.805) | (0.575) |
| Good Teacher Demeanor | -0.219 | -0.451 * | 0.496 | 0.723 | -0.664 ** | -1.031 *** |
| | (0.258) | (0.237) | (0.446) | (0.463) | (0.307) | (0.262) |
| Good Pedagogical Practice | -0.404 | -0.692 * | -0.181 | -0.493 | -0.471 | -0.825 ** |
| | (0.403) | (0.335) | (0.646) | (0.553) | (0.529) | (0.339) |
| Classroom Environment | -0.052 | -0.120 | 0.102 | -0.042 | -0.331 | -0.316 |
| | (0.187) | (0.163) | (0.238) | (0.180) | (0.249) | (0.198) |
| Availability of materials | 1.245 * | 1.361 ** | 0.881 | 0.430 | 1.567 * | 2.538 *** |
| | (0.698) | (0.623) | (1.250) | (1.054) | (0.795) | (0.617) |
| Use of materials | -0.309 | -0.061 | 0.005 | 0.354 | -0.486 | 0.064 |
| | (0.272) | (0.243) | (0.491) | (0.447) | (0.345) | (0.301) |
| **Stallings** | | | | | | |
| Share of time teaching | -0.158 | -0.388 | 0.207 | 0.079 | -0.428 | -0.317 |
| | (0.326) | (0.273) | (0.601) | (0.607) | (0.509) | (0.336) |
| Share of time lost | -0.759 | -0.601 | -1.283 | -1.613 | -0.469 | -0.604 |
| | (0.515) | (0.444) | (0.860) | (0.478) | (0.654) | (0.591) |
| Availability of materials | 0.470 | -0.045 | -0.350 | -1.223 | 1.602 | 0.829 |
| | (0.975) | (0.683) | (1.320) | (0.849) | (1.214) | (0.861) |
| Classroom environment | -0.128 | 0.034 | -0.152 | -0.047 | -0.184 | -0.128 |
| | (0.225) | (0.203) | (0.333) | (0.276) | (0.290) | (0.219) |
| **CLASS** | | | | | | |
| Emotional Support | 0.120 | 0.158 *** | -0.028 | 0.201 * | 0.304 *** | 0.025 |
| | (0.099) | (0.059) | (0.139) | (0.102) | (0.111) | (0.093) |
| Classroom Organisation | 0.155 | 0.190 | 0.214 | 0.246 | 0.142 | 0.183 |
| | (0.131) | (0.115) | (0.242) | (0.186) | (0.142) | (0.123) |
| Instructional Support | 0.147 | 0.159 *** | 0.029 | 0.217 ** | 0.281 ** | 0.085 |
| | (0.095) | (0.060) | (0.137) | (0.098) | (0.112) | (0.093) |
| Student Engagement | 0.093 | 0.121 ** | 0.001 | 0.205 ** | 0.205 ** | -0.002 |
| | (0.073) | (0.049) | (0.118) | (0.097) | (0.077) | (0.080) |
| Controls | NO | YES | NO | YES | NO | YES |

continued…

**Appendix 1 Table 9 continued: Student test scores regressed on classroom observation variables (level 1), variable-by-variable**

| | | | | | | |
|---|---|---|---|---|---|---|
| **Teach** | | | | | | |
| Share of time teaching | 0.462 | 0.106 | 1.129 ** | 0.651 | -0.534 | -0.464 |
| | (0.382) | (0.332) | (0.423) | (0.402) | (0.536) | (0.415) |
| Share of time lost | -0.623 * | -0.497 | -0.179 | 0.462 | -0.531 | -0.507 |
| | (0.349) | (0.394) | (1.844) | (1.638) | (0.346) | (0.409) |
| Classroom culture | 0.293 | 0.202 | 0.113 | 0.037 | 0.484 | 0.327 |
| | (0.209) | (0.189) | (0.313) | (0.292) | (0.305) | (0.246) |
| Instruction | -0.207 | 0.042 | -0.587 ** | -0.081 | 0.434 | 0.436 * |
| | (0.207) | (0.171) | (0.264) | (0.271) | (0.335) | (0.243) |
| Socioemotional Skills | 0.569 *** | 0.439 *** | 0.578 ** | 0.713 *** | 0.602 *** | 0.309 * |
| | (0.144) | (0.150) | (0.277) | (0.259) | (0.194) | (0.172) |
| Teacher Test Score | -0.017 | 0.032 | 0.214 *** | 0.188 *** | -0.433 | -0.353 |
| | (0.112) | (0.084) | (0.064) | (0.056) | (0.297) | (0.298) |
| Controls | NO | YES | NO | YES | NO | YES |

Note: Each coefficient and its associated standard error correspond to a separate regression of student test score on each variable. Standard errors clustered at the teacher level in parentheses. ***, **, and * indicate significance at the 1, 5, and 10