Teach Classroom Observation Tool

Background Paper

Measuring Teaching Practices at Scale

Results from the Development and Validation of the Teach Classroom Observation Tool

Ezequiel Molina
Syeda Farwa Fatima
Andrew Ho
Carolina Melo Hurtado
Tracy Wilichowskii
Adelle Pushparatnam

November 2018
Abstract

What goes on inside the classroom is central to student learning. Despite its importance, low and middle-income countries rarely measure teaching practices, in part due to a lack of access to adequate classroom observation tools and the high transaction costs associated with administering them. Teach, a new, open-source classroom observation tool for primary classrooms, was developed to capture the quantity and quality of teaching practices in these settings with a simple, easy-to-administer tool. This paper validates the use of Teach scores for system diagnostics by providing four types of evidence. First, it provides evidence that the practices included in the tool have a clear conceptual underpinning. Second, almost 90 percent of local observers in Mozambique, Pakistan, the Philippines, and Uruguay were highly accurate using Teach after a four-day training. Third, using data from 845 classrooms in Pakistan, the paper shows that Teach scores are internally consistent, present moderate to high inter-rater reliability in the field (.75 intraclass correlation coefficient), and provide substantial information that allows to differentiate teachers, even those with similar but not equal scores. Finally, teachers who display effective practices, as measured by Teach, are associated with students who achieve higher learning outcomes.

This paper is a product of the the Education. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/research. The authors may be contacted at molina@worldbank.org.
Measuring Teaching Practices at Scale

Results from the Development and Validation of the *Teach* Classroom Observation Tool*

Ezequiel Molina,a Syeda Farwa Fatima,b Andrew Ho,c Carolina Melo Hurtado,d Tracy Wilichowskia, and Adelle Pushparatnam,f

**JEL Classification:** I20; O15

**Keywords:** Education; Teacher Performance; Teacher Training; Education Policy and Planning; Public Service Delivery

---

*a This study was made possible by the World Bank’s Systems Systems Approach for Better Education (SABER) Trust Fund, which is supported by the United Kingdom’s Department for International Development (DFID) and Australia’s Department of Foreign Affairs and Trade (DFAT). We are grateful to the many researchers, survey experts, and observers who supported the data collection effort and to all of the participating schools in Punjab, Pakistan. Moreover, we thank the World Bank’s Punjab SABER SD team, led by Koen Martijn Geven, for helping coordinate the study. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors and do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank, its affiliate organizations, the Executive Directors of the World Bank, or the governments they represent. The *Teach tool* and all its complementary resources can be found [here](#).

bThe World Bank, molina@worldbank.org (corresponding author); bUniversity of Pennsylvania, Graduate School of Education; cHarvard University, Graduate School of Education; dUniversidad de Los Andes (Chile); eThe World Bank; fThe World Bank.
1. Introduction

School enrollment has increased substantially over the past 25 years in low- and middle-income countries. Schooling, however, does not guarantee learning. A large share of children complete primary school lacking even basic reading, writing, and arithmetic skills (World Bank, 2018) — a state of affairs UNESCO dubbed the “global learning crisis” (UNESCO, 2013).

The learning crisis is, at its core, a teaching crisis (Bold et al., 2017). A growing body of research indicates teaching is the most important school-based determinant of student learning (Hanushek & Rivkin, 2010; Snilstveit et al., 2016). The difference between the impact of a weak and great teacher on student test scores has been estimated at 0.36 standard deviations (SDs) in Uganda (Buhl-Wiggers et al., 2017) and 0.54 SDs in Pakistan (Bau & Das, 2017), respectively, which is equivalent to more than two years of schooling (Evans & Yuan, 2017). Moreover, evidence suggests several consecutive years of effective teaching can offset the learning shortfalls of marginalized students (Hanushek & Rivkin, 2006; Hanushek & Rivkin, 2010; Nye et al., 2004) and significantly improve students’ long-term outcomes (Chetty et al., 2011; 2014a; 2014b).

Despite its importance, identifying effective teaching is not easy and rarely done in practice. For example, Strong et al. (2011) showed that even experienced education professionals struggle to distinguish between effective and ineffective teachers. Most education systems in low- and middle-income countries do not regularly monitor teaching practices, or process quality (Ladics et al., 2018). Process quality refers to the interactions between teachers and students in the classroom (LoCasale-Crouch et al., 2016). Instead, education systems often choose to monitor elements of structural quality. Structural quality, on the other hand, refers to discrete elements that are indirectly related to teaching and learning and are easily observed, such as class size, teachers’ qualifications, and teacher training (Pianta, 2015).

Elements of structural quality such as teachers’ years of formal education (Staiger & Rockoff, 2010; Rivkin et al., 2005), years of experience (beyond the first two) (Araujo et al., 2016; Bau & Das, 2017; Rockoff, 2004), and entry exam performance (Cruz-Aguayo et al., 2017), only explain a small fraction of the variation in student learning and weakly predict process quality (Burchinal et al., 2002). In contrast, process quality has been shown to explain a larger share of student learning (Dobbie & Fryer, 2013; Hamre, 2014; Muijs et al., 2014). While the literature is far from a consensus on what share of the variation in student learning
can be explained by teaching practices (Burchinal, 2018; Leyva et al., 2015), there are several studies that highlight its importance for low- and middle-income countries, especially at the kindergarten level. For example, a study in Ecuador found a one SD increase in teaching practices is associated with a 0.18 SD increase in learning outcomes (Araujo et al., 2016). In Ghana, teaching practices account for a 0.07 to 0.17 SD increase in student learning outcomes (Wolf et al., 2018) – similar results are found in Chile (Leyva et al., 2015). Further, there is evidence that improvements in teaching practices lead to positive effects on student learning. For instance, a meta-analysis of over 60 coaching programs found those designed to improve teaching practices (0.58 SD), also resulted in increased learning outcomes (0.15 SD) (Kraft et al., 2018).

Even when education systems attempt to capture teaching practices, most tools used in low- and middle-income countries fall short on several accounts, as they: (i) measure either the quantity or quality of teaching practices; (ii) do not explicitly focus on teachers’ efforts to develop students’ socioemotional skills; (iii) use tools designed for other contexts, which may include irrelevant items or fail to include important ones; and (iv) use tools that are neither evidence-based nor meet basic reliability criteria.

The Teach classroom observational tool was developed in response to these concerns and to foster the measurement of teaching practices in low- and middle-income countries. Teach measures teacher practices at a primary school level and is intended to be used as a system diagnostic and monitoring tool and for professional development. As a diagnostic and monitoring tool at the system level, Teach helps governments identify bottlenecks in service delivery, monitor the effectiveness of their policies, and focus efforts to improve teacher practices. As a professional development tool, Teach can be used to identify individual teachers’ strengths and weaknesses and coach teachers to improve their practice.

In this paper, we made two contributions to the literature. The first contribution is primarily a methodological one. To the best of our knowledge, this is the first paper to use item response theory (IRT) for a classroom observation tool to go beyond reliability and factor analysis by assessing the information each element provides as it relates to teachers’ latent ability. The second contribution is empirical. This study validates the use of Teach scores for system diagnostics and monitoring with data from Punjab, Pakistan.

This paper is organized as follows. Section 2 presents the framework to validate the Teach scores for system diagnosis. Section 3 provides evidence on the tool’s content and cognition. Section 4 provides evidence of coherence and correlation. Section 5 concludes with a brief discussion of our findings.
2. Framework

We discuss five sources of validity evidence as outlined by *The Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). The central premise of validation is that the tool itself is not validated; rather, the scores from the tool are validated for a particular purpose (Kane, 2013), which applies to a given population of observers and users. This purpose could be for system diagnostics, teacher selection, or teacher evaluation. Sources of evidence are described by Ho (2018) using a “5 Cs” mnemonic: *content, cognition, coherence, and correlation.*

1. **Content:** There is theoretical and substantive evidence that *Teach’s* areas, elements, and behaviors measure teaching practices in low- and middle-income countries.

2. **Cognition:** There is evidence that Teach items are interpreted accurately by raters and aligned with the content.

3. **Coherence:** There is evidence that *Teach* is internally consistent and produces precise scores that can differentiate between teachers with similar, but not equal, ability. Scores do not vary substantially when two different observers view the same lesson.

4. **Correlation:** There is evidence that the *Teach* score is related to other metrics that have been found in the literature to be related to teaching practices (concurrent) and with student learning (predictive).

---

1 Validation frameworks ask for five types of evidence, including evidence that using *Teach* achieves its intended purpose in producing positive student learning outcomes. Ho (2018) describes this as a fifth “C”, Consequences. Development papers like this one should provide a theory of action for positive consequences but typically defer empirical evidence of positive consequences until after longitudinal outcomes accumulate.
3. Content and Cognition

3.1. Content

This section provides the rationale behind including both measures of teachers’ time on task and quality of teaching practices (Figure 1); the theoretical and empirical evidence behind Teach’s areas, elements, and behaviors; and how Teach overcomes the challenges that arise when applying existing classroom observation tools in low- and middle-income countries.

Figure 1: Teach Framework

3.1.1. Quantity and Quality

This subsection includes a discussion of how Teach measures quality and quantity of instruction and provides the evidence and ‘content validity’ for the Time on Task component.

As we mentioned briefly in the introduction, most tools used in low- and middle-income countries either capture the quantity or quality of teaching practices; however, they rarely capture both. For example, Stallings (Stallings, 1976), a commonly used low inference tool in low- and middle-income countries, employs a series of snapshots to determine – among others – whether students are on task and the amount of time the teacher spends teaching. The Classroom Assessment Scoring System (CLASS) (Pianta et al., 2008), a high inference tool that is less commonly used in low- and middle-income countries, examines the quality of teacher-student interactions. Although both tools are predictive of student learning outcomes

---

2 This reflects the final version of Teach, including the changes made as a result of this study detailed in Section 5.
(Bruns & Luque, 2014; Leyva et al., 2015), elements of what happens in the classroom are inevitably lost as neither tool captures quantity and quality. This problem is even more acute in low- and middle-income countries, which are often characterized by high absence rates and low instructional time (Bold et al., 2017; World Development Report, 2018).

The Teach classroom observation tool addresses this shortfall, as it measures the quantity and quality of teaching. Although the tool is primarily high inference, the Time on Task component includes a simplified version of the Stallings tool to capture whether teachers provide a learning activity and students are on task.

Research indicates effective teachers maximize the amount of time students spend on learning (Wharton-McDonald et al., 1998; Stronge, 2018). In fact, time lost in classroom instruction is associated with behavior issues and poorer student academic outcomes (Bruns & Luque, 2014; Dobbie & Fryer, 2013; Lavy, 2010; 2015). Using data from seven Latin American countries, Bruns and Luque (2014) showed that teachers from schools ranked in the top 25th percentile of student learning spent, on average, 80 percent of their time on task, as compared to teachers in the bottom 75th percentile, who, on average, spent only 30 percent of their time on task.

Like time on learning, student engagement is another important predictor of learning (Castillo, 2017). In the same study, Bruns and Luque (2014) found when students are on task and engaged, they learn significantly more than when they are distracted or off task.

3.1.2. What Are Effective Teaching Practices?

This subsection provides the evidence and ‘content validity’ for the teaching practices captured in Teach.3

The Quality of Teaching Practices component is organized into three areas: Classroom Culture, Instruction, and Socioemotional Skills. These areas have nine corresponding elements that point to 28 behaviors. The behaviors are characterized as low, medium, or high, based on the evidence collected during the observation. These behavior scores are then translated into a 5-point element scale that quantifies teaching practices, as captured in a series of two, 15-minute lesson observations.

Classroom Culture refers to a jointly-shared set of beliefs, attitudes, and behaviors by the teacher and students. Teachers who create a positive environment where students feel

---

3 This section is a summary of Molina et al. (2018a), which provides additional evidence on Teach ‘content validity’.
supported in their learning and encouraged to meet high academic and behavioral standards can have long-lasting, positive effects on students’ academic success (Burnett, 2002; Cornelius-White, 2007; Hamre & Pianta, 2006; OECD, 2009; Pianta, Hamre, & Stuhlman, 2003; Spilt et al., 2012). Teacher support to students can also reduce student internalizing (e.g., anxiety, depression) and externalizing (e.g., aggression) and enhance self-control (Grigg et al., 2016; Merritt et al., 2012).

Relatedly, teachers’ who are consistent and positive in establishing expectations not only help students reach their academic potential, but also support students’ development of positive behavior, social skills, and self-control within a safe environment (Jones, Bouffard & Weissbourd, 2013; OECD, 2009). Thus, another aspect of positive classroom culture requires teachers to prevent behavior problems and intervene when disruptive behaviors occur because such behaviors interfere with students’ learning and development (Stronge et al., 2007).

*Instruction* is essential for student learning (Carver & Klahr, 2001). Decades of research point to a few key features present in virtually all definitions of effective instruction. Effective teachers clearly deliver content in a way that entices students; they engage students in varied activities to promote thinking, assess students’ understanding, and offer feedback to students (Brophy, 1986; 1999; Porter & Brophy, 1988; Leyva et al., 2015). In fact, teachers who demonstrate these behaviors, compared to activities where the teacher is less involved, produce as much as a half of a standard deviation gain in student achievement (Hattie, 2009). Further, instructional support has been shown to be particularly beneficial for children with the lowest levels of academic abilities (Rimm-Kaufman et al., 2009).

Teachers who adapt their teaching strategies to meet the needs of their students can help them reach their potential. For instance, randomized intervention experiments conducted in India indicate that teaching that is tailored to students’ baseline level in mathematics has been found to improve children’s overall math scores by half a standard deviation point, with effects lasting after a year post-program conclusion (Banerjee et al., 2006). Similar effects of targeting teacher instruction and curriculum to students’ initial achievement level were found to be effective for Kenyan children as well, as this is thought to reduce the heterogeneity in the classroom learning environment (Duflo et al., 2011) and in Ghana, where significant improvements were found on closing children’s achievement gaps in numeracy and literacy skills after an in-school intervention (Duflo & Kiessel, 2017).

*Socioemotional Skills* development also plays an important role in academic achievement (Korpershoek et al., 2016). Despite a commonly held notion that there exists an artificial duality between the development of academic skills versus socioemotional skills,
effective classroom environments produce rigorous academic experiences in a socially-supportive classroom environment, thus promoting both academic and socioemotional development (Kochenderfer-Ladd & Ladd, 2015; Lee & Smith, 1999). In a recent meta-analytic study assessing the efficacy of social and emotional programs in kindergarten children through high-school students (N = 270,034), students experiencing programs designed to enhance socioemotional skills showed an 11-percentile-point increase in academic achievement (Durlak et al., 2011). Such programs showed positive consequences for improving student achievement and social and emotional skills, even beyond the length of the program (from 6 months up to 18 years after receiving programs) (Taylor et al., 2017).

Despite the importance of developing students’ socioemotional skills, few low- and middle-income countries measure them (Ladics et al., 2018). To address this gap, Teach measures how teachers support student Autonomy, which implicates students’ cognitive regulation skills, Perseverance, which exercises students’ emotional processes and cognitive regulation, and Social and Collaborative Skills, which requires students’ emotional processes and interpersonal skills. To our knowledge, this is the only classroom observation tool to include an area specifically designed to measure the ways teachers cultivate these skills.

Effective teachers foster autonomy in the classroom by creating opportunities for students to take ownership of their learning by building instruction around their interests, preferences, and choices (Evans & Boucher, 2015; Katz & Assor, 2007). If teachers use choice carefully and in a way that matches students’ interests and needs, students are more motivated and engaged, spend more time learning in ways that they prefer, can exercise their ability to assert their own opinion, and show better academic, behavioral and socioemotional outcomes (Fredricks et al., 2004; Jang et al., 2016; Katz & Assor, 2007; Reeve, 2006; 2009).

Learning requires effort; as such, failures and frustrations are inevitable. Thus, teachers need to encourage students to persevere through learning challenges by helping them understand their abilities and knowledge can be developed. This involves providing them with strategies for developing such abilities and knowledge and reassuring them that setbacks are integral parts of learning (Dweck, 1999; 2002; 2013). Teachers should also encourage students to set learning goals for themselves, and to persevere in their efforts to reach these goals (Duckworth et al., 2007). Teachers can have longstanding influence on their students’ perseverance, as demonstrated by one study in which sixth graders from Estonia reported on their teachers’ emotional support in their first three years of schooling. Students with the highest task persistence had teachers who score high on emotional support and low in psychological control in first grade (Kikas & Tang, 2018).
A teacher with a positive attitude toward students’ challenges can have a positive influence on student motivation and achievement. For instance, Zentall and Morris (2010) examined student responses to various scenarios illustrating failed behavior. Ultimately, they found that when most of the praise students received was non-generic (e.g., “you did a good job drawing”), students reported feeling happy about the scenarios, suggesting the emergence of a mastery orientation toward learning.

Finally, academic learning is an intensely social experience. Positive interactions with peers of the same age contribute to students’ academic, psychosocial, behavioral, and emotional well-being. These peer interactions take on increasing importance as children proceed through development (Parker & Asher, 1993; Hartup, 2009). Through peer relationships and experiences, children establish their concepts of trust, practice critical social skills, develop a sense of their own identity, and develop perceptions of other people and the world with lasting effects into later life.

The promotion of student collaboration in the classroom has benefits for children’s socioemotional development as well as their academic performance. For instance, in a study of Bruneian students, Kani and Shahrill (2015) found that where teachers assigned students to work in pairs to think-aloud and solve a set of math problems, improvements were observed in students’ problem-solving strategies and their understanding of the problem. Further, when paired with peers who are working at a slightly higher level of knowledge, scaffolding can occur; that is, the less-skilled peers’ memory recall and use of learning strategies improve while also increasing the more-skilled peers’ self-esteem (Manion & Alexander, 1997). This is also consistent with Wharton-MacDonald, Pressley, and Hampston’s (1998) study, which found that the most effective teachers with the highest performing students tended to encourage instructional groupings, where students would be encouraged to read or write with a partner during some part of the lesson or work in small-group activities cooperatively. Together this suggests that collaborations between students and peers, when structured well, can be conducive to positive learning outcomes for both parties.

3.1.3. Designed for Low- and Middle-Income Countries
This subsection provides evidence associated with issues that arise when applying existing instruments in low- and middle-income countries and how these issues have been addressed by Teach.
Teacher practices are commonly measured in low- and middle-income countries with either low inference tools developed locally or by high inference tools developed for the U.S. classrooms. The issue with locally developed low-inference tools is that they do not meet minimum reliability standards. This is evidenced by lax trainings, no exams to certify observers who understood the tool, and no studies computing the inter-rater reliability in the field (Ladics et al., 2018). Although high-inference tools, such the CLASS and Framework for Teaching (FFT), capture more nuance than low-inference tools, their use in low- and middle-income countries is also problematic for several reasons.

Because high-inference tools were developed for use in U.S. classrooms, users in low- and middle-income countries must adapt them for their context – however, there are no clear protocols to do this (Wolf et al., 2018). Without this adaptation, some of the tools’ items may not measure the same latent ability as others. For example, a recent study in Chile found that negative climate, which is a measure of negativity in the classroom (i.e. sarcasm, disrespect, anger, yelling), as captured by the CLASS, is not related to the emotional support domain (Leyva, 2015). Comparable results are found in Ecuador (Araujo et al., 2016). Moreover, preliminary results from the application of the CLASS in Tanzanian classrooms reveal that behavior management, as captured by the CLASS, does not seem to measure the same latent variable as the other CLASS constructs, displaying a negative correlation with the other constructs (Trako et al., 2018).

Teach was designed to address these concerns. The team first reviewed the literature on effective teaching practices in low- and middle-income countries (Molina et al., 2018a). These findings were then incorporated in the tool’s design. The tool was then revised based on feedback from over 20 education experts and tested in more than 10 low- and middle-income countries. This process led to an inclusion of new elements that were not captured in the original version of the tool, such as gender bias, and exclusion and revision of others.

Furthermore, the Teach team created two additional mechanisms to adapt the tool for use in low- and middle-income countries. First, local videos are used in trainings. Master coding with local videos ensures that Teach’s elements and behaviors are contextualized and anchored in the local setting. For example, although a Teach behavior states, “the teacher should treat all students respectfully,” evidence of what constitutes “respect” will vary – local, master coded videos can capture this nuance in a way that international videos cannot. Second, the tool is modular, meaning it allows users to add customized elements based on the
local curriculum and standards. This feature was piloted in Uruguay, where the local assessment agency worked closely with the *Teach* team to develop two new elements.

Another issue is that high-inference tools designed for the high-income countries do not provide ample granularity to differentiate between low performers. This has the effect of artificially bunching most teachers in low- and middle-income countries at the lower end of the scale. For example, most teachers (50% to 90%) who were observed using the CLASS in Afghanistan scored between a 1 and 2 on each construct in the Instruction domain (Molina et al., 2018b). Afghanistan is not an extreme case of this, as comparable results were found in Chile, where teachers scored between 1 and 3.8 on the Instruction domain, with a mean of 1.75 (Leyva et al., 2015). *Teach*, on the other hand, is designed to capture more granularity and differentiate among poor performers.4

Another challenge is that most high-income country tools are proprietary, costly, and difficult to implement (Bruns et al., 2016; Wolf et al., 2018). *Teach*, on the other hand, is freely available online and includes a suite of complementary materials. Aside from the manual and observation form available in English, French, Portuguese, and Spanish, *Teach* includes a complementary toolkit that helps users conduct the training with a detailed script and training guide, collect data using a data collection app available in several languages, and clean and analyze data with automatized programs — including assessing the validity of *Teach* scores. A template report to help communicate the results is also available.

3.2. Cognition

This section outlines evidence that *Teach* items are interpreted accurately by raters and aligned with the content as described in the previous section. This is achieved through simple descriptions, definitions, and examples in the manual. After 4 days of training, local observers go from being highly unreliable at the beginning of the training to reaching high scoring accuracy and interpreting the items appropriately by the end of the training. For example, in Pakistan after two days of the training, a mock exam reveals that only half of observers where scoring accurately. However, by the end of the training, 96% of them were able to pass the certification exam. This involves coding three videos and scoring at least 8 of the 10 elements in each video not more than one-point distance from the master codes.

---

4 See Section 4.2.3 for evidence of differentiation using the *Teach* tool.
Ladics and colleagues (2018) show most low-inference tools used in low- and middle-income countries rarely provide explanations or examples on the content they measure. For example, observers are asked to code whether the teacher provides “feedback” to students without an example or explanation of what “feedback” entails. In addition to ambiguous content, there is rarely accuracy criteria for observers to follow, and inter-rater reliability estimates are rarely conducted on data from the field.

High-inference tools, on the other hand, present different challenges as they were not designed for use in low- and middle-income countries. Evidence from the Measures of Effective Teaching (MET) study in the U.S. indicates that to obtain a 77% passage rate on the reliability exam (after two attempts), highly educated and experienced observers are needed to code high-inference tools. On average, all observers who participated in the study held a bachelor’s degree, two-thirds held a master’s degree, and about 7% held a Ph.D. Moreover, over three-fourths of observers had six or more years of teaching experience (Kane & Staiger, 2012).

Wolf and colleagues (2018) and Bruns and colleagues (2016) allude to the complexity of these tools and discuss potential difficulties in using them as a regular monitoring tool in low- and middle-income countries. To become certified on these tools, observers must undergo an extensive training to reliably code the nuances of high levels of practice. However, in most low- and middle-income countries, observers will never face those high levels of practice, as evidenced from the CLASS applications in Chile and Afghanistan (Leyva et al., 2015; Molina et al., 2018b). The use of U.S. videos for high-inference tool training is also problematic, as it does not adequately prepare observers for the challenges they will face while coding in low- and middle-income countries (Wolf et al., 2018).

To contextualize a tool for use in low- and middle-income countries, the Teach team designed a tool that is easy to follow, is concisely written, and includes specific examples. The length of the observation was determined based on rigorous evidence that shows that precise scores could be achieved in a 45-minute observation but also from numerous 15-minute observations (Ho & Kane, 2013). The observation form and electronic data collection application were created to minimize mistakes in data entry. In addition, the

---

5 As part of a different study (Trako et al., 2018) that used the CLASS for coding Tanzanian classrooms and recruited Tanzanian expats and foreigners who had previously lived in Tanzania to take the CLASS training, all of which had at least a master’s degree. Of the eight participants, only five passed the CLASS reliability exam.
training extends from the standard two days (for U.S. high-inference tools) to four days, which allows for additional time to practice coding additional videos, and conduct a field visit where participants participate in live classroom observations.

Before observers can code using Teach, they are required to pass a certification exam that involves scoring three videos within 1-point of the master codes at least 80% of the time. Analyses of the training data from Mozambique, Pakistan, the Philippines, and Uruguay indicate that of 145 participants, almost 90% (130), passed the exam. The lowest passage rate was in Mozambique, with 34 of 46 participants passing. The highest passage rate was in Uruguay, with 21 of 21 participants passing. In all four countries, local observers conducted the observations. These observers had a comparable level of education to the average citizen in their country and had no previous experience conducting classroom observations.

After the piloting, the tool underwent additional changes based on the feedback from training participants and data from the certification exam. For example, the critical thinking element was difficult for observers to understand and code consistently. To address this shortfall, the Teach team created a comprehensive table with examples for each subject and quality range. This was designed to aid the observers in processing and interpreting the information from the observation and manual.

4. Coherence and Correlation

4.1 Methods

4.1.1. Participants and Setting
We use data from the SABER Service Delivery (SABER SD) survey — which collects information with the goal of standardizing measures of student learning, teacher and school management quality, infrastructure and learning material, and student preparedness. The SABER SD survey grew out of a concern for poor learning outcomes, as evidenced by student tests and service delivery shortfalls. The survey builds upon the SABER (Systems Approach for Better Education Results) and SDI (Service Delivery Indicators) surveys (Bold et al., 2017).

For this study, we use a representative sample of 845 primary schools, consisting of 3,600 teachers and 19,000 students in Punjab, Pakistan (2018). In each school, one teacher from a randomly selected grade 4 classroom was observed using Teach for one, 20-minute

---

6 See Geven (2018) for details of the sample.
segment. In contrast to other applications of Teach, one, 20-minute segment was prioritized rather than two separate observations, as classes in Punjab last less than 30 minutes (in practice). The sampling frame was formed by primary schools with at least one fourth-grade class. The samples were designed to provide representative estimates of teacher effort, knowledge, and skills in public primary schools. The sample is further disaggregated by urban and rural localities and public and private schools.

The surveys collected a broad set of data on schools, teachers, and students. The information was largely collected via direct observation rather than from respondent reports. Data collection efforts, such as visual inspections of grade 4 classrooms and the school premises, administration of teacher and student tests, and physical verification of teacher presence by unannounced visits, were used to collect and verify the quality of the data.

The students and teachers in this chosen sample were also administered a content knowledge test as part of the study. The student test assesses grade 4 knowledge in English, Mathematics, and Urdu. The Mathematics test assessed students’ knowledge of number operations, measurement, geometry, Algebra, and data analysis. The English and Urdu test(s) assessed students’ knowledge of the alphabet, word recognition, word construction, grammar, vocabulary, sentence construction, and reading comprehension. This protocol and questionnaire were previously administered in Punjab; the scores from the test have been found to differentiate students of similar, but non-equal, abilities (Andrabi et al., 2007).

In contrast to assessments that simply require teachers to take an exam, our approach required them to mark (or “grade”) mock student tests in language and mathematics. This test simultaneously covered the same items as the student assessment. This method has two potential advantages: first, it aims to assess teachers in a way that is consistent with their regular teaching activities—namely, marking student work; second, by using a different mode of assessment for teachers, it distinguishes them as professionals. Previous versions of these instruments have been used in Sub-Saharan Africa, Afghanistan, and the Lao People’s Democratic Republic, which provide evidence that these measures are working as expected and are correlated with student outcomes (Bold et al., 2018; Molina et al., 2018b).

4.1.2. Procedures

The version of Teach used in Punjab has only minor differences with the one presented above. Section 5 discusses the rationale for the changes. The tool applied in Punjab had three areas: Classroom Culture, Instruction and Socioemotional Skills and 10
corresponding elements, which pointed to 27 behaviors. Like the current version of Teach, the behaviors are characterized as low, medium, or high, based on the evidence collected during the observation. These behavior scores are translated into a 5-point scale that quantifies teaching practices as captured in one 20-minute observation (as discussed above). After the 20-minute segment ends, the observer spends 15 minutes scoring the segment. Observations were conducted in person by two trained observers, who scored each segment independently. For each selected school, one grade 4 classroom was randomly selected to be observed during their mathematics or language class. For the analysis of the overall score and areas, we randomly selected one of the observers in the classroom rather than averaging the results from each observer, to obtain integer values.

To conduct the study, local observers were recruited through the RCons survey firm to collect data. None of them reported having had previous experience with classroom observation tools. Observers participated in a four-day training that required them to practice coding using recorded videos, participate in a live field visit, and pass a certification exam. The exam required them to code three 20-minute classroom observation segments in accordance with the manual’s rubric. After watching the 20-minute segment, observers were given 15 minutes to score the video. To pass the exam, they must be accurate within one of the master codes in eight of the 10 elements for each segment. Of the 53 observers who were trained and took the exam, 51 passed the exam. Table 1 displays the agreements between observers and master codes developed by Teach experts. The percentage of agreement within one point is 87%+ for all elements (except Feedback at 75%).

<table>
<thead>
<tr>
<th>Element</th>
<th>Exact Agreement</th>
<th>Within ± 1 Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supportive Learning Environment</td>
<td>42%</td>
<td>87%</td>
</tr>
<tr>
<td>Positive Behavioral Expectations</td>
<td>56%</td>
<td>98%</td>
</tr>
<tr>
<td>Opportunities to Learn</td>
<td>73%</td>
<td>99%</td>
</tr>
<tr>
<td>Lesson Facilitation</td>
<td>39%</td>
<td>90%</td>
</tr>
<tr>
<td>Checks for Understanding</td>
<td>55%</td>
<td>92%</td>
</tr>
<tr>
<td>Feedback</td>
<td>61%</td>
<td>75%</td>
</tr>
<tr>
<td>Critical Thinking</td>
<td>88%</td>
<td>100%</td>
</tr>
<tr>
<td>Autonomy</td>
<td>43%</td>
<td>100%</td>
</tr>
<tr>
<td>Perseverance</td>
<td>58%</td>
<td>98%</td>
</tr>
<tr>
<td>Social and Collaborative Skills</td>
<td>92%</td>
<td>99%</td>
</tr>
</tbody>
</table>

Note: The training sample includes observers who passed the Teach certification exam. To pass the exam participants need to be at most within one point of the master code in at least 8 out of the 10 elements in each video.
4.1.3. Data Analytic Approach

The analytical strategy for this study follows a four-step plan. Below we present the methods to be used and later the results.

4.1.3.1 Descriptive Analysis

To understand the data generating process and identify either floor or ceiling effect distributions, descriptive statistics are obtained for each of the 10 elements. Inter-item correlations are also computed to examine associations between the different elements of the *Teach* scale. Finally, inter-rater reliability was computed from field data to assess whether observers were reliable during the fieldwork. All classrooms were visited by two observers, which allows for estimation of measurement error due to raters.

4.1.3.2 Exploratory and Confirmatory Factor Analysis

Exploratory factor analysis (EFA) is conducted to explore the possible underlying factor structure of *Teach* elements, without imposing a preconceived structure. This allows for various combinations of the 10 elements to assess whether the data indicate one or multiple underlying qualities of teaching practices. Factor structures that contain one to three factors, using varimax and promax rotations, are examined. An ideal factor structure is chosen based on two criteria 1) the factor structure has at least three salient items per factor where loadings ≥ .4 indicate salience (Hair, et al., 1998; Tabachnick & Fidell, 2007) and 2) the factor structure makes theoretical sense in terms of parsimonious coverage of the data and compatibility with leading research in the area (Fabrigar et al., 1999).

Confirmatory factor analysis (CFA) verifies the factor structure of *Teach* elements. It allows us to test if the hypothesized relationship between *Teach* elements and the underlying latent construct exists. It does so using several indicators of the adequacy of model fit to the data. These goodness-of-fit tests are used to determine the appropriateness of the model. The chi-square test indicates the amount of difference between expected and observed covariance matrices; a chi-square value close to zero indicates little difference. The Comparative Fit Index (CFI) and Tucker Lewis Index (TLI) are equal to the discrepancy function adjusted for sample size. CFI and TLI range from 0 to 1 with a value of 0.90 or greater indicating an acceptable fit (Hu & Bentler, 1999). Root Mean Square Error of Approximation (RMSEA) is related to the residual in the model. It ranges from 0 to 1 with an RMSEA value of 0.06 or less indicating an acceptable fit (Hu & Bentler, 1999).
4.1.3.3 Item Response Theory

Descriptive statistics for item statistics (e.g., the percentage of teachers who score high on “fostering perseverance”) depend on the population being tested. IRT allows us to describe item characteristics that are expected to remain constant, regardless of the respondent population or the other items that accompany it, so long as the proposed IRT model fits the data. These item parameters include information and location. Information describes how well each element from Teach can distinguish among teachers with similar teaching ability, and location describes where on the Teach scale the particular element provides information to differentiate among teachers’ latent ability. An IRT graded response model for polytomously scored items is used to estimate information and location parameters for each element (Samejima, 1969).

Each element has location parameters equal to the number of response options minus one. For example, an element with five possible response options (1 to 5) will have four location parameters; therefore, every element in Teach has four location parameters (except for Critical Thinking and Perseverance since they have distributions that lie on a scale of 1-4).

These parameters are then used to derive a useful graph for evaluating the elements: item information functions. Item information functions present the overall level of distinction between teachers for each element. For a given teacher ability level, elements with higher peaks provide more precise information to differentiate teachers.

Summing up individual item information functions at each level of ability gives the overall test information curve. This shows how well Teach does holistically. In other words, it reveals at which range of teacher abilities Teach is providing the most information about differences among nearby examinees.

4.1.3.4 Concurrent and Predictive Associations

Concurrent associations provide evidence about the linear relationships between Teach and related outcomes measured at the same time. We examine the Teach score and individual factor scores for expected relationships with other related outcome measures, such

---

7 Item response theory (Lord, 1980), like factor analysis, is a method to estimate a respondent’s underlying ability/latent trait based on their answers to a series of items. IRT differs from factor analytic approaches by estimating and using information between item scale points (1 and 2 vs. 2 and 3, etc.) rather than assuming that these successive distances are equal. In our case, IRT provides estimates of teacher practices as captured by the Teach tool, which is based on the pattern of 1-5 scores.
as teacher characteristics, teachers’ subject and pedagogical content knowledge, and classroom facilities.

Predictive associations provide evidence about the extent to which Teach predicts scores on some criterion measure, such as student outcomes. Therefore, we first look at a simple correlation between standardized student assessment score and Teach score. We then repeat the analysis with a set of student controls (gender, age), teacher controls (gender, age, teaching experience, level of education), and classroom and school controls (class size, ownership status).

4.2 Results

4.2.1 Descriptive Analysis

The distributions of teacher practices for each of the 10 elements and 27 behaviors are displayed in Table 2. Overall, the distributions behave as expected, with higher scores on Classroom Culture and lower scores on Instruction and Socioemotional Skills. It is relevant to highlight two elements: Opportunities to Learn and Feedback. In the case of the former, we notice a ceiling effect, with most teachers providing most students with a learning activity most of the time. In the case of the latter, which is the only element with one behavior, we see observers tended to score the behavior as a low (1), medium (3), or high (5), which led to few instances of a 2 or 4.

Table 2: Overall Areas and Elements Means (SD) and Distribution of Teach Scores (N=845)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>[1-2]</th>
<th>[2-3]</th>
<th>[3-4]</th>
<th>[4-5]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overall Teach Score</strong></td>
<td>2.68</td>
<td>0.53</td>
<td>12%</td>
<td>63%</td>
<td>24%</td>
<td>1%</td>
</tr>
<tr>
<td><strong>Areas</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classroom Culture</td>
<td>3.87</td>
<td>0.63</td>
<td>2%</td>
<td>11%</td>
<td>52%</td>
<td>35%</td>
</tr>
<tr>
<td>Instruction</td>
<td>2.27</td>
<td>0.71</td>
<td>46%</td>
<td>42%</td>
<td>10%</td>
<td>2%</td>
</tr>
<tr>
<td>Socioemotional Skills</td>
<td>2.04</td>
<td>0.65</td>
<td>62%</td>
<td>33%</td>
<td>6%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td><strong>Elements</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supportive Learning</td>
<td>3.87</td>
<td>0.85</td>
<td>2%</td>
<td>4%</td>
<td>21%</td>
<td>51%</td>
</tr>
<tr>
<td>Environment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive Behavioral</td>
<td>3.28</td>
<td>0.86</td>
<td>4%</td>
<td>11%</td>
<td>42%</td>
<td>39%</td>
</tr>
<tr>
<td>Expectations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Opportunities to Learn</td>
<td>4.45</td>
<td>0.93</td>
<td>3%</td>
<td>3%</td>
<td>7%</td>
<td>22%</td>
</tr>
<tr>
<td>Lesson Facilitation</td>
<td>2.90</td>
<td>0.93</td>
<td>8%</td>
<td>24%</td>
<td>41%</td>
<td>26%</td>
</tr>
<tr>
<td>Checks for Understanding</td>
<td>2.55</td>
<td>0.97</td>
<td>17%</td>
<td>28%</td>
<td>40%</td>
<td>14%</td>
</tr>
<tr>
<td>Feedback</td>
<td>2.06</td>
<td>1.23</td>
<td>50%</td>
<td>13%</td>
<td>24%</td>
<td>9%</td>
</tr>
</tbody>
</table>
Critical Thinking 1.57 0.81 61% 25% 11% 3% 0%
Autonomy 2.47 0.97 19% 30% 38% 12% 1%
Perseverance 1.92 0.84 35% 42% 19% 4% 0%
Social and Collaborative Skills 1.73 0.95 54% 25% 14% 5% 1%

Note: In this Table, we present the mean, standard deviation, and distribution of the overall Teach score, Teach’s areas, and Teach’s elements. For the overall score and areas, we group the distribution in intervals, as we do not have integer values as in the case of the elements. We computed this by randomly selecting one of the observers in the classroom rather than averaging the result from each observer.

Figure 2: Histogram of Teach Overall Score and Elements

Note: In this Figure we present the distribution of the overall Teach score and each Teach’s element.

Table 3 presents the descriptive statistics and inter-item correlations of the 10 elements with means ranging from 1.57 to 4.45 and inter-item correlations ranging from .01 to .43.

Table 3: Teach Inter-element Correlations (N=845)

<table>
<thead>
<tr>
<th></th>
<th>SLE</th>
<th>PBE</th>
<th>OL</th>
<th>LF</th>
<th>F</th>
<th>A</th>
<th>P</th>
<th>SCS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supportive Learning</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Environment</td>
<td></td>
<td>0.32*</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive Behavioral Expectations</td>
<td>0.22*</td>
<td>0.24*</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Opportunities to Learn</td>
<td>0.28*</td>
<td>0.31*</td>
<td>0.21*</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lesson Facilitation</td>
<td>0.22*</td>
<td>0.23*</td>
<td>0.23*</td>
<td>0.43*</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 4 presents the inter-rater reliability estimates, which confirm the findings from the training. Observers maintain and even slightly improve their reliability while in the field. Exact agreement ranges from 54% (Supportive Learning Environment and Positive Behavioral Expectations) to 79% (Opportunities to Learn). Within one agreement is above 90% for all elements, except Feedback (87%). Finally, the ICC ranges from 0.53 for Supportive Learning Environment and Positive Behavioral Expectations to 0.81 for Opportunities to Learn. This is generally understood as moderate (0.50-0.75) and good (0.75-0.9) values for ICCs (Koo & Li, 2016).

Table 4: Teach Fieldwork Inter-Rater Reliability

<table>
<thead>
<tr>
<th>Element</th>
<th>Exact Agreement</th>
<th>Within ± 0.5 Agreement</th>
<th>Within ± 1 Agreement</th>
<th>ICC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supportive Learning Environment</td>
<td>54%</td>
<td>95%</td>
<td>95%</td>
<td>0.53</td>
</tr>
<tr>
<td>Positive Behavioral Expectations</td>
<td>54%</td>
<td>94%</td>
<td>94%</td>
<td>0.53</td>
</tr>
<tr>
<td>Opportunities to Learn</td>
<td>79%</td>
<td>97%</td>
<td>97%</td>
<td>0.81</td>
</tr>
<tr>
<td>Lesson Facilitation</td>
<td>57%</td>
<td>97%</td>
<td>97%</td>
<td>0.69</td>
</tr>
<tr>
<td>Checks for Understanding</td>
<td>57%</td>
<td>94%</td>
<td>94%</td>
<td>0.64</td>
</tr>
<tr>
<td>Feedback</td>
<td>62%</td>
<td>87%</td>
<td>87%</td>
<td>0.67</td>
</tr>
<tr>
<td>Critical Thinking</td>
<td>67%</td>
<td>95%</td>
<td>95%</td>
<td>0.6</td>
</tr>
<tr>
<td>Autonomy</td>
<td>60%</td>
<td>95%</td>
<td>95%</td>
<td>0.7</td>
</tr>
<tr>
<td>Perseverance</td>
<td>61%</td>
<td>94%</td>
<td>94%</td>
<td>0.58</td>
</tr>
<tr>
<td>Social and Collaborative Skills</td>
<td>64%</td>
<td>93%</td>
<td>93%</td>
<td>0.63</td>
</tr>
</tbody>
</table>

Note: In this Table, we present the proportion of exact and adjacent agreements and the ICC for the field sample. To compute this type of reliability, we collected data from two independent observers for each classroom (N=845) and measured the agreements between their scores. Exact Agreement measures the proportion of observations where observers were in exact agreement for each element. Within ± x Agreement measures the proportion of observations where observers were either in exact agreement or within x points of one another. The ICC ranges from 0 to 1 and is a widely used measure of inter-rater reliability that accounts for mean differences between the raters as a component of variability.

4.2.2 Exploratory and Confirmatory Factor Analysis
A one-factor solution is chosen based on our criteria. All elements have salient factor loadings ≥ .4 (except for Social and Collaborative Skills). Figure 3 is a scree plot from this analysis that indicates the first dimension accounts for substantially more variation than subsequent dimensions. We conclude that a single dimension underlies item responses, which we assert based on content and cognition evidence as the quality of teaching practices. CFA also confirms this one-factor solution. The model has good statistical fit with $\chi^2 (35) = 145.91$, CFI=.92, TLI=.90 and RMSEA=.06, verifying a single common factor explains the relationships among the element scores.

We also test a three-factor solution and find good fit statistics with $\chi^2 (df) = 98.51$ (32), CFI=.95, TLI=.94 and RMSEA=.05. We interpret the likelihood ratio test, showing a three-factor solution has a better statistical fit than a one-factor solution, as evidence supporting the three underlying subdomains, largely due to our large sample size enabling detection of small differences. However, taken together, results in exploratory and confirmatory factor analysis suggest a one-factor solution is parsimonious and consistent with the intended use of the Teach score. Therefore, a one-factor solution is chosen. The resulting average score also demonstrates high internal consistency (Cronbach’s alpha ($\alpha$) = .77).

Figure 3: Teach Dimensionality

Note: Scree plot of eigenvalues show the variation accounted for by each element (out of 10). This is estimated from a 1-factor EFA (N=845).
4.2.3 Item Response Theory

An IRT graded response model is fit to the full sample to estimate location and information parameters. Table 5 reports the estimated location and information parameters for each element.

**Table 5: Item Information and Location Parameter Estimates Based on a Graded Response Model (N=845)**

<table>
<thead>
<tr>
<th>Element</th>
<th>Information Parameter</th>
<th>b1</th>
<th>b2</th>
<th>b3</th>
<th>b4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supportive Learning Environment</td>
<td>0.72</td>
<td>-6.01</td>
<td>-4.16</td>
<td>-1.48</td>
<td>2</td>
</tr>
<tr>
<td>Positive Behavioral Expectations</td>
<td>1.02</td>
<td>-3.5</td>
<td>-1.99</td>
<td>0.33</td>
<td>3.59</td>
</tr>
<tr>
<td>Opportunities to Learn</td>
<td>0.77</td>
<td>-5.06</td>
<td>-4.05</td>
<td>-2.79</td>
<td>-0.9</td>
</tr>
<tr>
<td>Lesson Facilitation</td>
<td>1.56</td>
<td>-2.12</td>
<td>-0.71</td>
<td>0.87</td>
<td>3.25</td>
</tr>
<tr>
<td>Checks for Understanding</td>
<td>1.69</td>
<td>-1.36</td>
<td>-0.18</td>
<td>1.45</td>
<td>3.49</td>
</tr>
<tr>
<td>Feedback</td>
<td>1.54</td>
<td>-0.03</td>
<td>0.46</td>
<td>1.65</td>
<td>2.59</td>
</tr>
<tr>
<td>Critical Thinking</td>
<td>1.2</td>
<td>0.43</td>
<td>1.79</td>
<td>3.4</td>
<td></td>
</tr>
<tr>
<td>Autonomy</td>
<td>1.44</td>
<td>-1.38</td>
<td>-0.05</td>
<td>1.75</td>
<td>3.78</td>
</tr>
<tr>
<td>Perseverance</td>
<td>1.33</td>
<td>-0.64</td>
<td>1.16</td>
<td>2.91</td>
<td></td>
</tr>
<tr>
<td>Social and Collaborative Skills</td>
<td>0.55</td>
<td>0.33</td>
<td>2.58</td>
<td>5.12</td>
<td>8.91</td>
</tr>
</tbody>
</table>

Note: In this Table, all items have a 1-5 score distribution range and thus 4 location parameters. In the case of Critical Thinking and Perseverance, the actual distributions lay between 1-4 and thus 3 location parameters.

The information parameters range from 0.55 for Social and Collaborative Skills to 1.69 for Checks for Understanding. These results indicate that Teach items effectively distinguish between teachers, even those with similar levels of latent ability. Higher values indicate the element provides more information. In this case, it means Check for Understanding provides more information about the quality of teaching practices than Social and Collaborative Skills or Opportunities to Learn.

Location parameters describe where on the Teach scale the information is located, for each successive score point (between 0 and 1, 1 and 2, etc.). The lower location estimates (b1) distinguish between lower-performing teachers, while the higher location estimates (b4) distinguish between higher-performing teachers. Higher absolute values indicate the element differentiation is greater at that location on the scale. For example, the Supportive Learning Environment thresholds are better at distinguishing between lower-performing teachers, while Social and Collaborative Skills thresholds are better at distinguishing between higher-performing teachers.
Figure 4 presents item information functions for each element. Checks for Understanding and Lesson Facilitation have the highest peaks and therefore provide the most information across teachers of varying abilities. Opportunities to Learn and Social and Collaborative Skills have flatter curves at the bottom and therefore distinguish the least across teachers of different abilities.

**Figure 4**: Item Information Functions for each Teach Element Estimated from a Graded Response Model (N=845)

Using information from all the elements, the test information function is drawn (Figure 5). Teach provides balanced information across nearly the entirety of the scale (-2 to +4 SD) and is maximized (i.e., can differentiate more) for teachers with average levels of ability. Since Social and Collaborative Skills provides the least information, we compare the test information function with and without it.⁸ This element provides little information to the right tail but some information to the left tail, which increases the precision of the Teach score.

---

⁸ Results are available upon request.
4.2.4 Concurrent and Predictive Associations

Table 6 displays correlations between Teach scores (overall and IRT rescaled) and teacher and classroom characteristics. The Teach score is positively and significantly associated with teachers’ subject and pedagogical content knowledge, education level (having completed college or university), having more than two years of experience, and contractual status (temporary), ranging from |.05| to |.22|. These findings are consistent with recent research that shows positive associations between teachers’ content knowledge and first two years of experience with teacher value added (Bau & Das, 2017). Further, we see a positive relationship between Teach scores and the presence of basic classroom facilities for the teacher and students (i.e. blackboard, chalk, desks, chairs, books, stationary), which ranged from .03 to .16. This is consistent with earlier research that shows physical classroom resources are associated with student behavior management and the nature of interactions students experience in the classroom (Bennell & Akyeampong, 2007; Wolf et al., 2018). In summary, we observe positive but small to moderate correlations, which is consistent with earlier literature on the topic.
Table 6: Concurrent Correlations

<table>
<thead>
<tr>
<th>Teacher Characteristics</th>
<th>Teach Score</th>
<th>IRT rescaled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject Content Knowledge</td>
<td>.06*</td>
<td>.08*</td>
</tr>
<tr>
<td>Pedagogical Content Knowledge</td>
<td>.05*</td>
<td>.07*</td>
</tr>
<tr>
<td>Teacher education</td>
<td>.14*</td>
<td>.17*</td>
</tr>
<tr>
<td>(Completed college/university)</td>
<td>.08*</td>
<td>.09*</td>
</tr>
<tr>
<td>Temporary Teachers</td>
<td>-.20*</td>
<td>-.22*</td>
</tr>
<tr>
<td>Classroom Facilities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classroom Hygiene</td>
<td>.06*</td>
<td>.07*</td>
</tr>
<tr>
<td>Chalk and Blackboard</td>
<td>.07*</td>
<td>.08*</td>
</tr>
<tr>
<td>Teacher: Desk</td>
<td>.16*</td>
<td>.19*</td>
</tr>
<tr>
<td>Teacher: Chair</td>
<td>.10*</td>
<td>.11*</td>
</tr>
<tr>
<td>Student: Chair and Desk</td>
<td>.03*</td>
<td>.05*</td>
</tr>
<tr>
<td>Student: Textbook</td>
<td>.07*</td>
<td>.06*</td>
</tr>
<tr>
<td>Student: Pen and Pencil</td>
<td>.08*</td>
<td>.07*</td>
</tr>
<tr>
<td>Student: Exercise Book</td>
<td>.08*</td>
<td>.07*</td>
</tr>
</tbody>
</table>

Note: *p<0.05. We defined Classroom Hygiene equal to 1 if the classroom is extremely or reasonably clean as evaluated by the observer; Chalk and Blackboard are equal to 1 if they are both present; Teachers: Desk is equal to 1 if present; Teacher: Chair is equal to 1 if present; Student: Chair and Desk is equal to 1 if all students have them; Student: Textbook is equal to 1 if 80% or more students have them; Student: Pen and Pencil is equal to 1 if 80% or more students have them; Student: Exercise Book is equal to 1 if 80% or more students have them.

Table 7 presents regression estimates with a set of student, teacher, and school controls. A unit increase in Teach score is associated with .13-.24 (.07-.14 using IRT rescale score) standard deviation increase in student test score.

Table 7: Teach Associations with Student Learning

<table>
<thead>
<tr>
<th>Dep. Variable</th>
<th>(1) OLS Total Score</th>
<th>(5) OLS Total Score</th>
<th>(6) OLS Math Score</th>
<th>(7) OLS Urdu Score</th>
<th>(8) OLS English Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teach</td>
<td>0.14*** (.050)</td>
<td>0.21*** (.048)</td>
<td>0.15*** (.046)</td>
<td>0.13*** (.042)</td>
<td>0.24*** (.048)</td>
</tr>
<tr>
<td>Constant</td>
<td>-0.4*** (0.141)</td>
<td>-0.7*** (.211)</td>
<td>-0.7*** (.198)</td>
<td>-0.3 (.183)</td>
<td>-0.7*** (.214)</td>
</tr>
</tbody>
</table>

Observations 18,243 14,947 14,947 14,842 14,842
Adj. R-squared 0.005 .097 .061 .064 .097
Number of schools 845 845 845 845 845
Student Characteristics X X X X X
5 Discussion

We use a 4-step validation framework to provide evidence that Teach is a valid measure for practitioners to monitor teaching practices. This evidence is based on data collected from a large representative sample of schools in Punjab, Pakistan. We provide evidence that the measured content has a clear theoretical and empirical foundation (content), that the elements and behaviors are interpreted and used correctly by observers (cognition), that the components relate to one another as expected, that the elements are internally consistent and the score is reliably captured (coherence), and that the score is correlated with other measures, as expected by the literature (correlation). To the best of our knowledge, this is the first paper to use IRT to assess the information and location of each element in a classroom observation tool.

There are several new insights that our analysis reveals. The Teach elements provide substantial information on the quality of teaching practices. Not only that, but this information is distributed evenly throughout the scale, which allows us to use Teach to differentiate among teachers of similar but not equal ability. We can also better understand what information the tool is gathering and identify three key areas for improvement.

First, we noticed Opportunities to Learn has strong ceiling effects, a relatively low level of information, and overlap between the possible response curves. This indicates that for this element, teachers of different latent abilities obtain the same score. Second, since Feedback had only one behavior, it becomes a de facto three-point element. As a result, while Feedback does contain relatively high levels of information, its response curve overlaps, which hinders the extent to which teachers can be differentiated as they rarely score in the 2 or 4-point range. Finally, like Opportunities to Learn, Supportive Learning Environment also has strong ceiling effects; in fact, of these three elements, the least amount of information can be gathered from Supportive Learning Environment.

We also show Teach has a low to moderate correlation with other variables, as expected from the literature. Moreover, increases in Teach scores predict higher student
learning outcomes, even after controlling for the teacher, student, school, and classroom characteristics.

Our results have implications for Teach’s content and future research on classroom observation tools. Regarding the former, after this pilot concluded, Teach’s elements were revised based on insights from the data, training data, and feedback from observers and experts. Aside from revisions to the tool’s descriptions and examples, three main changes were made.

First, Opportunities to Learn is now referred to as Time on Learning, which is measured through a series of snapshots (Bruns & Luque, 2014; Stallings, 1976). This element now captures how teachers use classroom time and the proportion of student engagement during this time. Past applications of Stallings have shown these measures are associated with student learning outcomes (Bruns & Luque, 2014; Bruns et al., 2016). Second, Feedback now includes two, rather than one behavior: behavior one measures the extent to which the teacher helps clarify students’ misunderstandings and behavior two measures the extent to which the teacher identifies students’ successes. Moreover, we added one additional behavior to the Social and Collaborative Skills element to measure the extent to which the teacher promotes students’ collaboration. Lastly, we modified the gender bias behavior, within the Supportive Learning Environment element, to not only capture teachers’ biases, but also the extent to which they actively challenge gender stereotypes in the classroom (Molina et al., 2018a).

Regarding future research, there are several avenues to enhance our understanding of Teach. First, we should assess the differential item analysis by comparing the data from the application of Teach in different countries. Second, we should conduct a generalizability study to understand the reliability and precision of the tool for different uses. Third, we should conduct a study using longitudinal analysis to compare the precision and predictive validity of different classroom observation tools.

As we conclude, it is important to note the limitations of this study. First, the study is descriptive in nature. It uses cross-sectional data – meaning relationships between teacher practices and student outcomes are not causal. Second, the teaching practices are based on just one, 20-minute classroom observation. The measurement error inherent in these types of studies could also affect the magnitude and strength of the association with student outcomes. Third, evidence on coherence and correlation from this validation study is based on a sample of observers and schools in Punjab, Pakistan. While the results from this study provide strong confidence on the use of Teach in low- and middle-income countries to monitor teaching
practices additional studies will need to be conducted in other countries to further corroborate these findings.

References


Reeve, J. (2009). Why teachers adopt a controlling motivating style toward students and how they can become more autonomy supportive. *Educational Psychologist*, 44(3), 159-175.


