WPS8396

Policy Research Working Paper 8396

Understanding Effective Teaching Practices in Chinese Classrooms

Evidence from a Pilot Study of Primary and Junior Secondary Schools in Guangdong, China

Andrew Coflan
Andrew Ragatz
Amer Hasan
Yilin Pan

WORLD BANK GROUP
Education Global Practice
April 2018

Policy Research Working Paper 8396

Abstract

This study documents the results of a pilot study jointly undertaken by the World Bank and the Guangdong Department of Education to assess teaching practices in public primary and junior secondary schools using the Classroom Assessment Scoring System tool. The tool was used to conduct classroom observations on an illustrative sample of 36 teachers in three counties across Guangdong. The pilot tested whether such a tool could be used to measure the strengths and weaknesses of teaching practices in the Chinese context. It also informed how classroom observations can be more consistently applied in the province's Quality Assurance and Monitoring and Evaluation system. On average, teachers in this sample scored high on classroom organization, but lower on emotional support and instructional support. While there was substantial variation in performance across teachers, there was only modest variation by county, urban versus rural school location, teacher type, grade, and years of experience. Teachers who believed that students should be the focus of instruction (those who espouse student-centered learning) scored significantly higher across all domains than teachers who believed that teachers should be the focus of instruction (those who espouse teacher-centered learning). Together the results from this pilot provide insights into how teacher training can address the most critical gaps in teaching practices.

This paper is a product of the Education Global Practice. 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://econ.worldbank.org. The authors may be contacted at aragatz@worldbank.org.

Understanding Effective Teaching Practices in Chinese Classrooms: Evidence from a Pilot Study of Primary and Junior Secondary Schools in Guangdong, China

Andrew Coflan

Andrew Ragatz

Amer Hasan

Yilin Pan

JEL codes: I21, I28

Keywords: Teachers, classroom observation, teacher training

Acknowledgments

The authors gratefully acknowledge support from the Results in Education for All Children (REACH) Trust Fund. We would like to thank the Guangdong Department of Education as well as representatives from the counties of Dianbai, Lianjiang, and Wuhua for their assistance in collecting the data. Teachstone provided training to classroom observers and certified their proficiency. They also provided valuable follow-up advice during data analysis. Constructive inputs and feedback were received by Lorena Sernett, Sarah Hadden and Daniel LaCava for the final report. Any errors are our own. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

1 INTRODUCTION

China has made huge strides in human development in the past three decades. Going forward, a key government priority is ensuring equality of quantity and quality of education. As student learning is dependent on teacher quality, school systems must ensure that their teachers receive systematic and targeted training to support effective teaching practices.

A basic challenge in ensuring teacher quality is lack of agreement on how to consistently measure instructional quality and teaching practices in the classroom. In Guangdong, teachers are often observed by other local teachers, the department heads of their schools, or by school administrators. These observations, however, are more subject to personal teaching style than any type of official best practices, and are often described as a checklist focused on a small range of behaviors. It is therefore difficult, if not impossible, to measure teaching quality in any standardized or comparable way, or to address shared shortcomings in teaching quality by education officials.

The work described in this pilot study was undertaken with three objectives in mind. The first was to assess whether, and to what extent, an internationally validated measure of teaching practices can be applied to Chinese classrooms. The second was to gain an understanding of the process involved in using direct classroom observation as a means of assessing teaching practices. This was needed not only for this pilot study but also for longer-term monitoring and evaluation that will be conducted under a loan from the International Bank for Reconstruction and Development to the People's Republic of China for the proposed Guangdong Compulsory Education project. The third was to assess whether the information collected through classroom observation can be used to provide feedback to training providers, with a view to developing a results-based financing mechanism for improving teacher training.

Given these three objectives, our key research questions focus on the results of classroom observations and responses to teacher beliefs surveys.

1. Can such a tool be used to measure the strengths and weaknesses of teaching practices in the Chinese context, and if so, what do teaching practices in a selection of Guangdong schools look like?
2. Do these teaching practices vary across key characteristics such as geography, school type or teacher type?
3. What is the relationship between teacher beliefs and teaching practices?

This study applies the Classroom Assessment Scoring System (CLASS)—a classroom observation tool that measures 12 dimensions of teaching on a scale of 1–7.[1] In this study, information is also collected on teacher backgrounds and teacher beliefs.

1.1 Background

To decide on the appropriate instrument for this pilot study, three approaches to classroom observation were considered: the Stallings Method, the Third International Mathematics and Science Study (TIMSS) video study coding method, and the CLASS instrument (see section 3.1). A small-scale field test was conducted to assess the feasibility of the three approaches (from November to December 2015). Based on the findings of this field test and a review of the literature, the CLASS tool was selected for this study. This is backed up by the findings of the Measures of Effective Teaching (MET) project, a three-year study across seven US public school districts designed to determine how best to identify and promote great teaching. The research from the MET project identified teacher assessments with the use of observation tools as

[1] CLASS was developed by researchers at the University of Virginia; CLASS observation training is now administered and developed by Teachstone.

particularly effective when feedback was given to teachers. The MET project identified the CLASS instrument as being particularly effective in correlating teacher assessment results with student learning outcomes.

Box 1. Glossary

Backbone teacher – A senior teacher in the Chinese school system, recognized both for quality of teaching and contributions to their school. Backbone teachers receive additional teacher training and often take larger roles in mentoring other teachers.

CLASS—The CLassroom Assessment Scoring System, a method for observing teachers.

Cohen's Kappa—A measure of inter-rater reliability.

Dianbai—A county in Guangdong where the CLASS instrument was piloted.

Dimension—Individual indicator within the CLASS instrument. Measures a specific behavior, action, or relationship. See table 3.

Domain—An overarching category comprising several CLASS dimensions. There are three domains: Emotional Support, Classroom Organization, and Instructional Support.

Inter-rater Reliability (IRR)—A measure of agreement between pairs of raters observing the same lesson. Measured using a variety of statistical techniques.

Lianjiang—A county in Guangdong where the CLASS instrument was piloted.

Wuhua—A county in Guangdong where the CLASS instrument was piloted.

Zhongkao—A test taken by all Chinese middle-school students at the end of 9th grade. Results determine where they attend upper-secondary school.

This study focuses on assessing teaching practices in three "pilot counties"—Wuhua, Dianbai, and Lianjiang—representing high, medium, and low education performance respectively per the Guangdong Department of Education. These counties are among the 16 "project counties" identified for inclusion in the Guangdong Compulsory Education project.

Table 1 gives some economic and demographic context. Average gross domestic product (GDP) per capita of the 16 project counties is less than one-third of Guangdong's provincial average. Similarly, the project counties produce one-fifth of their GDP from agriculture, with almost three-quarters of the population living in rural areas. For Guangdong, less than 2 percent of GDP comes from agriculture, which is less than one-tenth the average for the project counties.

Table 1: Basic Data, Guangdong and Project Counties, 2014

Province / Counties	GDP per capita (Y)	GDP from agriculture (%)	Rural population (%)	Population
Guangdong	63,469	1.7	32.0	107,240,000
Wuhua	11,452	21.9	86.7	1,363,000
Dianbai	28,969	19.5	61.4	1,460,000
Lianjiang	24,006	25.9	79.2	1,485,000
Average of 16 project counties	18,957	20.1	74.3	1,231,024

Source: Project counties, Guangdong Department of Statistics 2015.

For this study, administrative data from Guangdong province were collected on county performance on the Zhongkao, a county-level test taken by all 9th grade students to determine where they will attend upper-secondary school. Data were collected from the 16 project counties as well as 20 comparator counties within Guangdong. Ten of the comparator counties were designated as impoverished by the province; 10 were not. The 16 project counties generally underperform the comparator counties. Of the 16 project counties, 13 scored in the bottom half of the 36-county sample. The average score for the 16 project counties was 6 percentage points lower than the average of the 20 comparator counties, and 10 percentage points lower than the non-impoverished counties.

Taken together, these factors indicate the need for a comprehensive teacher assessment and training system that can both address the difficulties in measuring effective teaching and help improve those areas in which teachers are falling behind.

1.2 Roadmap

Section 2 presents a short literature review, the CLASS instrument, the data, and collection methods. It also presents evidence of reliability of the data collected. Section 3 analyzes the results, looking at differences across CLASS domains and dimensions, and by pilot county and by each of the teachers observed. It then attempts to tease out strands in the relationship between teacher beliefs and teaching quality. Section 4 presents some conclusions and recommendations. There are four appendixes.[2]

2 Literature Review and Methodology

2.1 Literature review

Evans and Popova (2017) conducted a literature review of meta-analyses and found pedagogical interventions (e.g., Conn 2014; Kremer, Brannen, and Glennerster 2013) and teacher training (e.g., McEwan 2014) to be among the most effective interventions in improving student learning outcomes. But

[2] Appendix A examines the CLASS results in further detail. Appendix B goes into greater detail regarding the statistical validity of the CLASS instrument in the pilot counties. Appendix C proposes a framework for a results-based financing mechanism. Appendix D details the complete protocols for CLASS observations.

it is notoriously difficult to change teaching practices, and most interventions fail unless they are designed well and targeted properly.

China, like other countries, is working to improve its teacher training programs, investing millions of dollars in professional development. Yet a recent study by the Stanford Rural Education Action Program found that, while teachers in rural China gained knowledge from their training, teachers did not subsequently change their teaching behaviors, leading to no significant improvement in student outcomes (Lu et al. 2017). To address such challenges, one promising avenue is to have professional development programs accompanied by classroom observations and feedback. A special type of observation is one not done directly in the classroom, but on a deferred basis, based on records of practices that have sufficient detail to allow coding (or rating) of recorded behaviors, for instance with video recordings (Martínez-Rizo 2012).

There is growing interest in observing teaching practices, for formative feedback to teachers, teacher performance evaluation and incentives, program impact evaluation, and research on the determinants of student learning (Bruns 2016). Tools have evolved over time. Many early versions tended to take the form of checklists (whether a teacher used certain techniques, whether work of children was displayed on the wall, etc.). More detailed instruments, such as the "Stallings Classroom Snapshot" instrument, developed by Professor Stallings in the 1970s, captured low-inference measures on amounts of time spent on specific activities (e.g. proportion of classroom time spent on group work), frequencies (e.g. number of questions asked by the teacher), and other general observations (e.g. proportion of students engaged in the activity at specific points in the lesson). These forms of observation provided a useful picture of what was taking place in the classroom and allowed analysts to infer certain levels of quality in teaching-learning, but little was directly measured on the teaching-learning process.

A more recent generation of instruments aims to measure the quality of what takes place in the classroom. These instruments are based on current theories on the issues to observe and are more sophisticated in dealing with psychometric properties (Martínez-Rizo 2012). These instruments tend to use scales and be based on rubrics that define a range of quality from low to high. Many of these instruments have been developed in the United States. The most prominent of these are CLASS itself, developed by Robert Pianta and colleagues, and the Framework for Teaching, developed by Charlotte Danielson.

Some instruments were developed expressly for developing countries, such as Teacher Instructional Practices and Processes System (TIPSS); and for specific subjects, such as the Protocol for Teaching Language Arts Observation (PLATO), Mathematical Quality of Instruction (MQI), and Quality of Science Teaching (QST).

For this pilot study CLASS was chosen because it has proven to be one of the most valid and reliable lesson observation instruments and there is consistent evidence that its measures have a strong relationship with student learning outcomes.

2.2 Methodology

Data collection for this pilot study took place between December 2016 and January 2017. This study is based on a purposive sample of the three pilot counties. Each pilot county had four schools observed, and each school provided three teachers for observation. Among the four schools, two were elementary schools and two junior secondary schools, one of each rural ("out of township") and one of each urban ("in township") (table 2).

Classroom observations were conducted in primary and junior secondary schools. Raters assessed the 36 teachers of English, Mathematics, and Chinese in their regular classrooms. Thirty teachers were videotaped, some of who were also observed live, while six had live-only evaluations. The teachers were a mix of 4th and 8th grade teachers. At each school, county officials were asked to identify three types of teachers: those who were "new to teaching" (had less than three years of experience—but see endnote 4); those who had been designated "backbone teachers;" and those who had been identified as potential backbone teachers—

"backbone candidates". Backbone teachers are those identified by either the county Bureau of Education or the provincial Department of Education as excellent teachers; such teachers generally receive additional training beyond what the general teacher gets and are usually accorded additional responsibilities within their schools. Backbone candidates are those who have been identified as potential backbone teachers, but who have yet to go through the requisite training to qualify as full backbone teachers. Supplementing the classroom observations was a teacher beliefs survey and training history questionnaire given to all 36 observed teachers.

The breakdown of teachers (see table 2), while not large enough to form a representative sample, allows us to judge whether CLASS is useful in the Chinese context, as well as to examine teacher performance from several angles: teacher experience, overall academic performance, rural versus urban,[3] and by subject.[4]

Table 2: Breakdown of Teacher Observation Sample

	Primary (4th grade)			Junior Secondary (8th grade)			
	Urban (in township)	Rural (out of township)	Total primary	Urban (in township)	Rural (out of township)	Total Jr. Sec.	Total
Overall sample							
Total schools	**3**	**3**	**6**	**3**	**3**	**6**	**12**
Total teachers	**9**	**9**	**18**	**9**	**9**	**18**	**36**
New teacher	*3*	*3*	*6*	*3*	*3*	*6*	*12*
Backbone candidate	*3*	*3*	*6*	*3*	*3*	*6*	*12*
Backbone teacher	*3*	*3*	*6*	*3*	*3*	*6*	*12*
Sample within a single county							
Schools within a county	**1**	**1**	**2**	**1**	**1**	**2**	**4**
Teachers per school	**3**	**3**	**6**	**3**	**3**	**6**	**12**
New teacher	*1*	*1*	*2*	*1*	*1*	*2*	*4*
Backbone candidate	*1*	*1*	*2*	*1*	*1*	*2*	*4*
Backbone teacher	*1*	*1*	*2*	*1*	*1*	*2*	*4*
Sample by subject							
Math	3	3	**6**	3	3	**6**	**12**
Language (Chinese)	3	3	**6**	3	3	**6**	**12**
English	3	3	**6**	3	3	**6**	**12**

[3] While all schools observed are in Guangdong's poorer counties, the conditions in each of these counties vary depending on whether one is in the countryside or in the county seat. For this study, half of the schools observed are in county seats, while half are in the countryside. Despite this distinction between urban and rural in our data set, it is likely that *all* schools observed would be classified as rural relative to true urban centers such as Guangzhou, Shanghai, or Beijing.

[4] The only divergence between initially planned criteria and the final breakdown was the selection of "new" teachers. While new teachers were originally requested to have less than three years of experience, the final sample included "new" teachers with up to 19 years of experience. Follow-up discussions with the government suggest that they consider teachers with more than three to five years of experience who are not backbone candidates as "general" or "common" teachers, rather than new teachers.

2.2.1 Classroom observation training

To pilot the CLASS tool in Guangdong, a team of World Bank researchers and experts from Teachstone prepared a five-day training course, held in September 2016. The CLASS instrument and relevant training materials were translated into Chinese by the World Bank. Teachstone trainers traveled to China with the World Bank team to deliver the training. Thirty-three individuals participated in the training to become classroom raters. They were from the Guangdong Department of Education, as well as from selected normal universities and teaching colleges in the province. There was a mix of PhD candidates, university professors, department heads and deans, and education officials.

At the end of the training, the classroom raters had to pass a certification test to be able to participate in the data collection exercise. A passing score was 80 percent. Of the 33 recipients of training, 28 passed. An extract of classroom observation protocols, adapted from Teachstone materials, is in Box 2.

2.3 The CLASS instrument

A classroom observation tool to measure the quality of teacher–student interactions, CLASS is subject-agnostic and has been applied across countries and grades, from pre-primary through secondary school. It is based on developmental theory and research that demonstrates that interactions between teachers and students are the primary mechanism through which children learn. CLASS is one of the most widely researched lesson-observation instruments, with over 150 studies conducted. The efficacy of CLASS as a measure of instructional quality has been supported by studies involving thousands of classrooms and tens of thousands of students across age levels. Collectively, these studies indicate that classroom quality, as measured by the CLASS, predicts positive developmental and academic outcomes for children (e.g. Hamre et al. 2012).

While most of the research on CLASS has been conducted in the United States, an expanding body of research is international, and applicability of measurement systems at this level is an important step, especially when assessments are not focused on objective knowledge, such as how to solve a math equation, but rather on subjective measures of student–teacher interactions (e.g. Pakarinen et al. 2010; von Suchodoletz et al. 2014).

CLASS is designed around a student-centered learning environment, while Chinese classrooms have traditionally been focused on examining lessons in terms of curriculum and learning goals. The early childhood instrument for CLASS has been implemented in Guangdong and was shown to be valid (Hu et al. 2016), but has not yet been validated at the Upper Elementary level.

Issues regarding the applicability of CLASS to the Chinese context were raised both by raters and peer reviewers, who emphasized the differences between Chinese and U.S. classrooms and attitudes toward education (see section 2.4). These concerns are valid. As documented in the sections that follow, all statistical measures, both on inter-rater reliability and the outcomes of the measurement tool, are equivalent to U.S. comparator studies. This supports the usability of CLASS in the Chinese context.

The exact actions that CLASS observes vary by grade, but there are broad overlaps in the general categories of behaviors. This assessment used the Upper Elementary CLASS tool, which has 12 dimensions. Eleven dimensions are aggregated under three domains, while Student Engagement is its own separate category (table 3).

Box 2: Extract of Classroom Observation Protocols

Arrival at the School

Arrive at the school at least one hour before the scheduled video shooting. Late arrival can create difficulties in the preparations for the filming. The teacher will have a set time for his or her lesson and the set-up should not alter the lesson time.

First meet with school officials. You should never go directly to the teacher's classroom. Always go to the main office first and meet with the principal or the person who has been assigned as your official contact person.

Once in the Classroom

As soon as you get to the classroom where you will shoot the lesson, two factors will help you determine where to position the camera: information about what will happen during the lesson, and the physical arrangement of the room.

Ask the Teacher about the Lesson

Try to find out from the teacher about what will happen in the lesson. Often there will be little time for you to talk to the teacher because even though you arrive early, he or she might be busy teaching. However, if you have a chance, ask the teacher and find out as much as possible about the lesson.

Timing of Observations

Each coded section should be 20 minutes long. Rather than code the entire session, the class period should be divided into 20-minute sections, each to be graded separately.

Document the Teacher

During the lesson, teachers engage in a variety of activities. For example, they explain concepts and procedures, pose problems, assign tasks, ask questions, write information on the chalkboard, walk around the classroom, and assist individual students.

Because the main goal of this pilot is to study teaching practices, it is necessary that we thoroughly and carefully document the teacher's activities and behaviors during the lesson. Make sure that you capture what the teacher is doing, what he/she is saying, and what information he/she is presenting to the class.

Document the Students

Make sure that you capture what students are doing and saying during the whole-class interaction, when they are working in groups and on their own. Focus mainly on the activities and behaviors of the students who are interacting with the teacher, but turn to other students as well from time to time because students might be doing different things when the teacher is and is not with them. The goal is to sample student behavior so that what is portrayed in the videotape is representative of what happened in the lesson.

Document the Tasks

Normally the teacher presents the task to students clearly enough that students understand what they are supposed to do, and it is usually not hard to see in the video what the task is. This is not always the case, however. If the task is ambiguous or poorly described, many students will be uncertain how to proceed. In all cases, what we want to see on the video is the task that students are engaged in doing, whether it is what the teacher intended. Make sure you document how students are doing the assigned tasks.

Table 3: Description of Upper Elementary CLASS Domains, Dimensions, and Indicators			
Domain	**Dimension**	**Description**	**Indicator and Behavioral Markers**
Emotional Support	**Positive Climate**	Reflects the emotional connection between the teacher and students and among students.	**Relationships**—Physical proximity, shared positive affect, peer interactions, social conversation
			Positive affect—Smiling, laughter, enthusiasm
			Positive communications—Positive comments and expectations
			Respect—Cooperation, use of each other's names, listening to each other, warm/calm voice, respectful language
	Teacher Sensitivity	Encompasses the teacher's awareness of and responsiveness to students' academic and emotional needs.	**Awareness**—Anticipates problems, checks in with students, notices difficulties
			Responsiveness to academic and social/emotional needs and cues—Acknowledgement of emotions, individualized support, reassurance/assistance, re-engagement, timely response, adjusts pacing/wait time as needed
			Effectiveness in addressing problems—Student issues/questions resolved, follow up
			Student Comfort—Seek support and guidance, take risks, participate freely
	Regard for Student 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.	**Flexibility and student focus—**Follows student's lead, shows flexibility, encourages student ideas and opinions
			Support for autonomy and leadership—Allows choice, chances for leadership, gives students responsibility, relaxes structure for movement
			Connections to current life—connects content to student life, communicates usefulness
			Meaningful peer interactions—peer sharing and group work
Classroom Organization	**Behavior Management**	Encompasses the teacher's ability to provide clear behavioral expectations and use effective methods to prevent and redirect misbehavior.	**Clear expectations**—Consistent. explicit, students know what to do
			Proactive—Anticipation of problem behavior, low reactivity, attention to the positive, monitoring, proximity
			Effective redirection of misbehavior—Uses subtle cues to redirect, peer redirection and problem solving, little time lost, problems resolved
			Student behavior—Compliance with teacher, little aggression or defiance, meets expectations, absence of chaos
	Negative Climate	Reflects the overall level of expressed negativity in the classroom.	**Negative affect**—Irritability, anger, harsh voice, physical aggression, disconnected or escalating negativity
			Punitive control—Yelling, threats, physical control, harsh punishment
			Disrespect—Bullying, teasing, humiliation/sarcasm, exclusionary behavior, inflammatory/discriminatory/derogatory language or behavior
	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.	**Maximizing learning time**—Tasks provided, disruptions minimized, choice when finished, effective completion of managerial tasks
			Routines—Students know what to do, little wandering, clear instructions
			Transitions—Little wasted time,, redirection to task when necessary, time cues provided
			Preparation—Materials ready and accessible, knows lessons
		Focuses on the way in which the teacher maximizes	**Active facilitation**—Teacher interest, effective pacing, promoting involvement

Table 3: Description of Upper Elementary CLASS Domains, Dimensions, and Indicators			
Domain	**Dimension**	**Description**	**Indicator and Behavioral Markers**
	Instructional Learning Formats	students' interest, engagement, and ability to learn from lessons and activities	**Variety of modalities, strategies, and materials**—variety of modalities and strategies, variety of materials, interactive materials
			Effective engagement—Active participation, sustained attention
			Learning targets/organization—clear learning targets, previews, reorientation/summary statements, clear/well-organized presentation of information
Instructional Support	**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.	**Depth of understanding**—Multiple and varied perspectives, real world connections, emphasis on meaningful relationships among facts, skills, and concepts
			Communication of concepts and procedures—essential components identified, conditions for how and when to use the concept and/or procedure, multiple and varied examples, contrasting non-examples
			Background knowledge and misconceptions—attention to prior knowledge, explicit integration of new information, attention to misconceptions, students share knowledge and make connections
			Transmission of content knowledge and procedures—clear and accurate definitions, effective clarifications, effective rephrasing
			Opportunity for practice of procedures and skills—supervised practice, independent practice
	Quality of Feedback	Assesses the degree to which the teacher provides feedback that expands learning and understanding and encourages continued participation.	**Feedback loops**—Back and forth exchanges, persistence, follow-up questions
			Scaffolding—Assistance, hints, prompting completion and thought processes
			Building on student responses—Expansion, clarification, specific feedback
			Encouragement and affirmation—Recognition and affirmation of effort, encouragement of persistence
	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.	**Facilitation of higher-order thinking—**Students identify and investigate problems/questions, students examine, analyze, and/or interpret data, information, approaches, students construct alternatives/predict/hypothesize/brainstorm, students develop arguments, provide explanations
			Opportunities for novel application—Students apply previous knowledge/skills, presents cognitive challenges, open-ended tasks
			Metacognition—Students explain their own cognitive processes, students self-evaluate/reflect/plan, teacher models thinking about thinking
	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.	**Cumulative content-driven exchanges**—connection to content, depth of exchanges, exchanges that build on one another
			Distributed talk—Balance of student and teacher talk, student-initiated dialogues, majority of students, peer dialogues
			Facilitation strategies—Open-ended questions/statements, students respond, acknowledgement/repetition/extension, active listening, pause as needed to allow thinking and full expression
	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.	**Active engagement** Responding, asking questions, volunteering, sharing ideas, looking at the teacher, active listening, manipulating materials, lack of off-task behavior

Source: Upper Elementary CLASS manual. Used with permission of Teachstone.

Each dimension looks for the presence of indicators, each characterized by the presence of a variety of behavioral markers as expressed by the teacher and student, grouped in three ranges: a score of 1 or 2 (the

low range) describes few instances of the behavioral markers; a score of 3, 4 or 5 (the mid-range) describes intermittent or mixed instances of the behavioral markers; and a score of 6 or 7 (the high range) describes frequent and consistent instances of the behavioral markers (table 4). Quality and depth of indicators is also taken into account when judging score.

Table 4: CLASS Likert Scale 1–7

Low Range		Mid Range			High Range	
1	2	3	4	5	6	7
The low-range description fits the classroom/ teacher very well. All, or almost all, relevant indicators in the low range are present.	The low-range description mostly fits the classroom/ teacher but there are one or two indicators that are in the mid range.	The mid-range description mostly fits the classroom/ teacher but there are one or two indicators in the low range.	The mid-range description fits the classroom/ teacher very well. All, or almost all, relevant indicators in the mid-range are present.	The mid-range description mostly fits the classroom/ teacher but there are one or two indicators in the high range.	The high-range description mostly fits the classroom/ teacher but there are one or two indicators in the mid range.	The high-range description fits the classroom/teacher very well. All, or almost all, relevant indicators in the high range are present.

Source: Upper Elementary CLASS Manual, page 12. Used with permission of Teachstone.

Raters documented evidence of each indicator as they occurred within a 15–20-minute period, and then rated the frequency of the observed behaviors as in the low, mid, or high range. The behavioral markers are intended to capture overall quality of the interaction and are not intended to serve as a checklist. The notes corresponding to each score in table 4 are not meant to be comprehensive but rather abbreviated guides around which a rater was allowed to recall the frequency of such instances. Finally, dimension scores within specific domains were averaged to come to an average score for Emotional Support, Classroom Organization, and Instructional Support, respectively. The only dimension for which this is not the case is Negative Climate, where lower scores are better, as they indicate an absence of negativity.[5] For the purposes of calculating the Classroom Organization domain score, the Negative Climate score is reversed by subtracting it from 8, allowing the dimension to be averaged with others in the domain.

2.4 Inter-rater reliability

This section presents the results of a series of analyses to gauge whether the raters' results were reliable. These results are then compared against other CLASS studies, identifying strengths and weaknesses in the results from the pilot counties. The pilot study had 12 raters, forming 13 different pairs, making a total of 302 individual observations over the course of the study. Each rater had received and passed CLASS training 18 weeks before rating the observed teachers.

Three comparator studies, which took place in a U.S. context, are presented in table 5 to provide a reference for other implementations of the CLASS tool in upper elementary and junior secondary.

[5] For more guidance on computing CLASS scores, see pages 17-19, K-3 CLASS Manual.

Table 5: Three Comparator United States–based Studies

Study	Number of Teachers	Grades	Raters	Supporting Organization	Subjects	Period
Measures of Effective Teaching (MET)	1,333	4–9	Graduate Students at University of Virginia (UVA)	Bill & Melinda Gates Foundation	Math and English	2009
Understanding Teaching Quality in Algebra Study (UTQ-A)	82	6–12	Trained Raters at Educational Testing Service	UVA, W.T. Grant and Spencer Foundations	Math and English	2009–11
Secondary MyTeachingPartner Study (S-MTP)	78	7–12	Graduate Students at UVA	UVA	No subject specified	2007–09

Measuring agreement allows us to see where raters had an easier or harder time in agreeing on observed behaviors. It is also in line with CLASS protocols, as the same measure is used to evaluate whether a rater has passed CLASS training.

Table 6 shows the average percentage agreement between raters across the 12 CLASS dimensions. The raters in Guangdong had similar measures of agreement to those in the three other studies. Certain dimensions proved harder for the Chinese raters to assess similarly, such as Regard for Student Perspectives, Instructional Dialogue, Teacher Sensitivity, and Analysis and Inquiry, perhaps due to differing definitions of what comprises effective interactions for each of these dimensions in the context of a Chinese classroom. On other measurements, such as Productivity and Student Engagement, the Chinese raters agreed a higher percentage of the time than the raters in the other three studies.

Table 6: Agreement, Exact + Adjacent (%)

	Guangdong	S-MTP	MET	UTQ-A
Positive Climate	78.4	79.3	74.6	69.3
Teacher Sensitivity	58.4	73.8	72.6	65.0
Regard for Student Perspectives	64.8	78.6	67.9	70.2
Behavior Management	80.4	89.2	85.6	92.8
Productivity	91.5	84.3	82.4	87.2
Negative Climate	95.6	95.1	95.1	97.6
Instructional Learning Formats	72.0	80.3	78.2	73.9
Content Understanding	70.7	73.4	75.5	76.6
Analysis and Inquiry	66.4	73.7	71.5	84.3
Quality of Feedback	78.1	72.5	71.9	64.1
Instructional Dialogue	64.8	—	74.8	—
Student Engagement	83.1	81.2	78.4	82.0

Source: Guangdong observations; Teachstone.

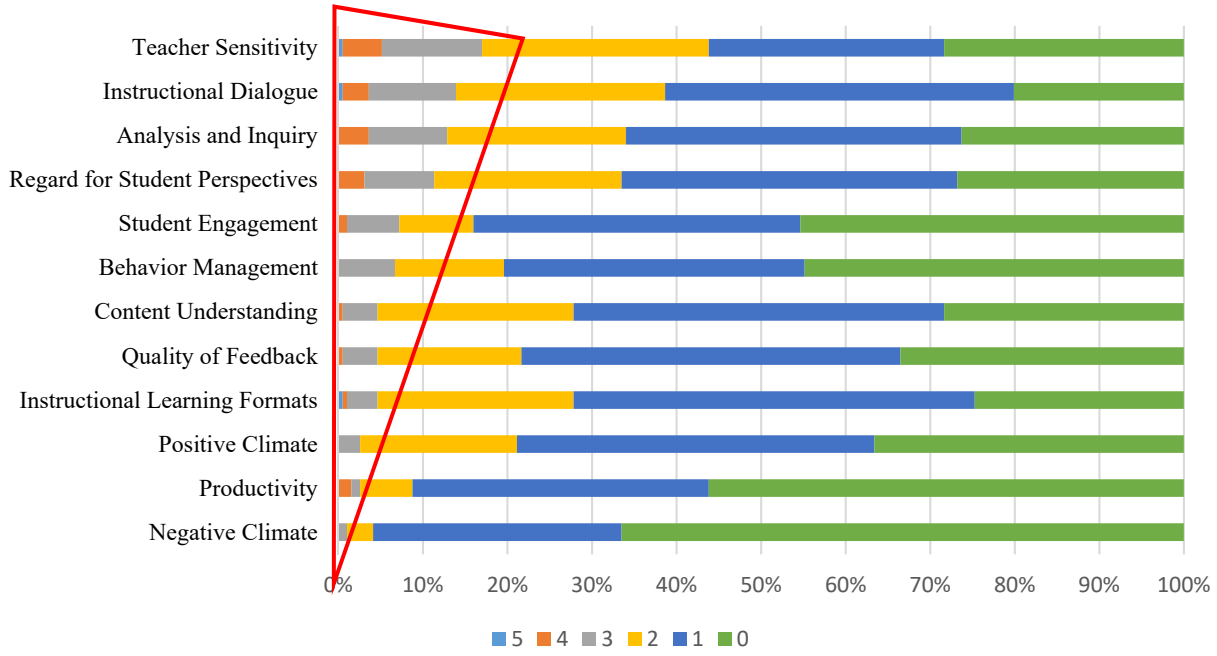
Note: — indicates that the dimension was not scored for that study. Adjacent indicates (±1) between the two raters. For example, if one rater scored a certain dimension a 4, while their partner rated it a 3, 4, or 5, that would count as being in agreement

An alternate method of considering differences in coding is to examine the spread between scores of the raters for the same lesson segment. For example, two raters may give a score of 5 on the dimension of

Teacher Sensitivity, in which case there is no spread between their scores. But if one rater scores 2 and the other 5, there is a spread of 3. A score of 3 or more indicates a large discrepancy in judgment.

Figure 1 shows that a few dimensions have a significant proportion of coding with a spread of 3 or more. For Teacher Sensitivity, Instructional Dialogue, Analysis and Inquiry, and Regard for Student Perspectives, at least 10 percent of discrepancies had a spread of 3 or more.

Figure 1: Extent of Coding Differences for Each Dimension



Source: Guangdong Pilot Study; authors' calculations.

A session was held with the raters in a June 2017 workshop to understand why these discrepancies arose. The raters indicated that the main factors were technical and cultural. For Analysis and Inquiry and Instructional Dialogue, the challenges appear to be mainly technical understanding of the dimensions and could be addressed through more intensive training. While the technical element also played a role for Teacher Sensitivity and Regard for Student Perspectives, there were important cultural elements to consider. These dimensions bring in new concepts that are not typically considered in the lesson observations taking place in Guangdong. The raters generally saw value in these dimensions, but proposed that the indicators be examined to consider how they might be modified to better reflect the Guangdong context. For further measures of observation reliability, including intraclass correlation coefficient and Cronbach's alpha analysis, see Appendix B.

3 RESULTS

3.1 Overall

The data suggest that teachers have strong Classroom Organization skills. The scores on Emotional Support and Instructional Support are lower. This pattern is in line with evidence collected in other settings around the world. Teachers were notably weaker in Emotional Support and Instructional Support than they were in Classroom Organization. Low Emotional Support and Low Instructional Support are the two most highly correlated domains. While not necessarily a causal relationship, a focus on either domain will likely lead to improvements in both.

Given the small sample sizes involved in this pilot study, readers are cautioned against overinterpreting the results. Still, the differences that do (and do not) exist all provide insight into the state of teaching among the 16 project counties in Guangdong and have useful lessons for further teacher training and observation.

The results presented in this section focus on the applicability of CLASS within the context of the proposed Guangdong Compulsory Education project. For an analysis of breakdowns by years of teaching experience, method of observations, and more detailed breakdowns by county, see Appendix A.

3.1.1 Across domains

Figure 2 shows the average scores across the three CLASS domains, separated by grade. Teachers observed in the three pilot counties scored highest in Classroom Organization. Elementary-school teachers scored higher than middle-school teachers across all three domains.

Figure 2: Domain Results, Guangdong Pilot Counties

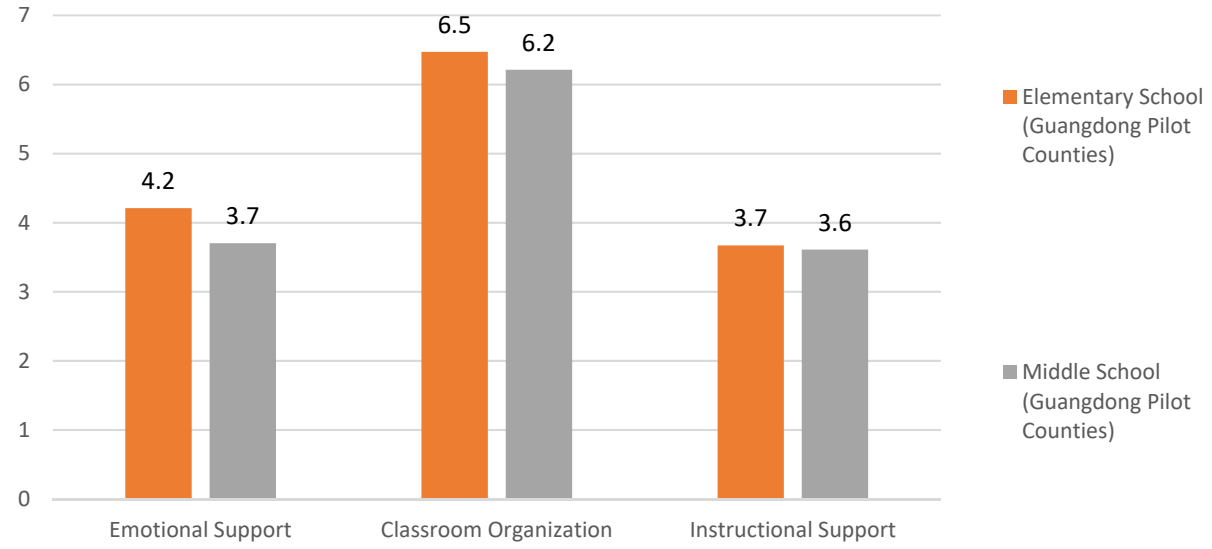
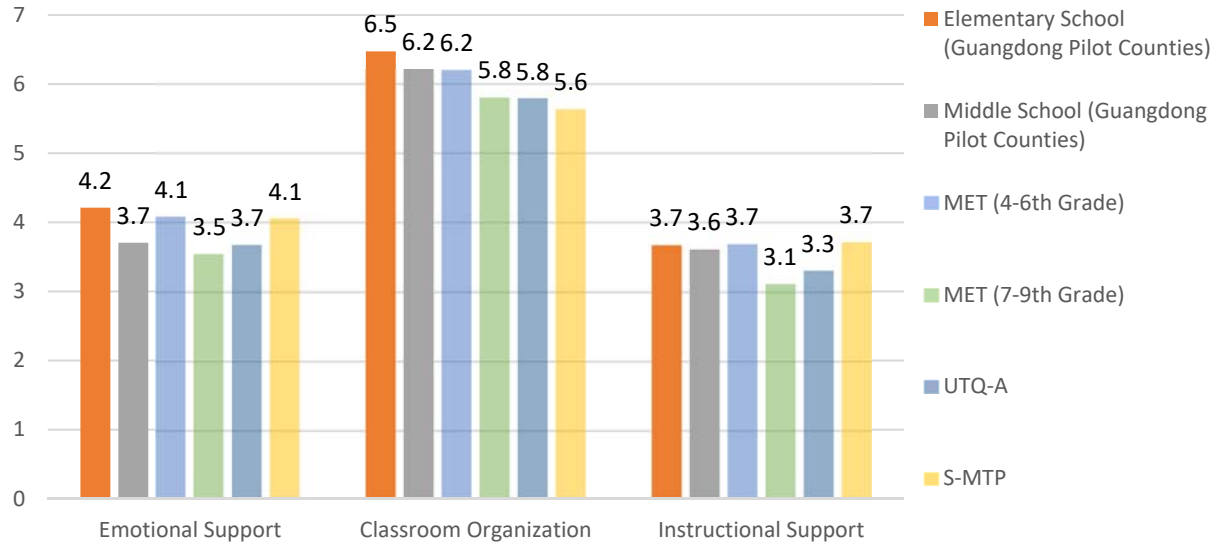


Figure 3 adds four comparator values from the three studies, all using the Upper Elementary CLASS observation tool. The three studies are all from the United States, however, and therefore provide a limited international comparison. They are the largest upper elementary CLASS studies available.

It shows that, first, both in the Guangdong and MET studies, lower-grade classrooms scored higher than middle-school classrooms across all three domains. While the effect is less pronounced in the Guangdong scores, it is nevertheless consistent throughout. Second, despite popular discussion on the differences in classroom culture and the role of the teacher between China and the United States, overall scores are quite similar among the four studies, particularly on Emotional Support and Instructional Support. The largest difference between the three pilot counties and the U.S. studies is in Classroom Organization. Finally, Instructional Support is the weakest domain across all studies. Further work must be done to determine whether the results obtained in Guangdong for the domain of Instructional Support are related to difficulty of measuring some of the domains (particularly Analysis and Inquiry and Instructional Dialogue) or to an overall weakness on the part of teachers in providing adequate, student-centered instructional support. The approach of this pilot study does not allow for this to be assessed.

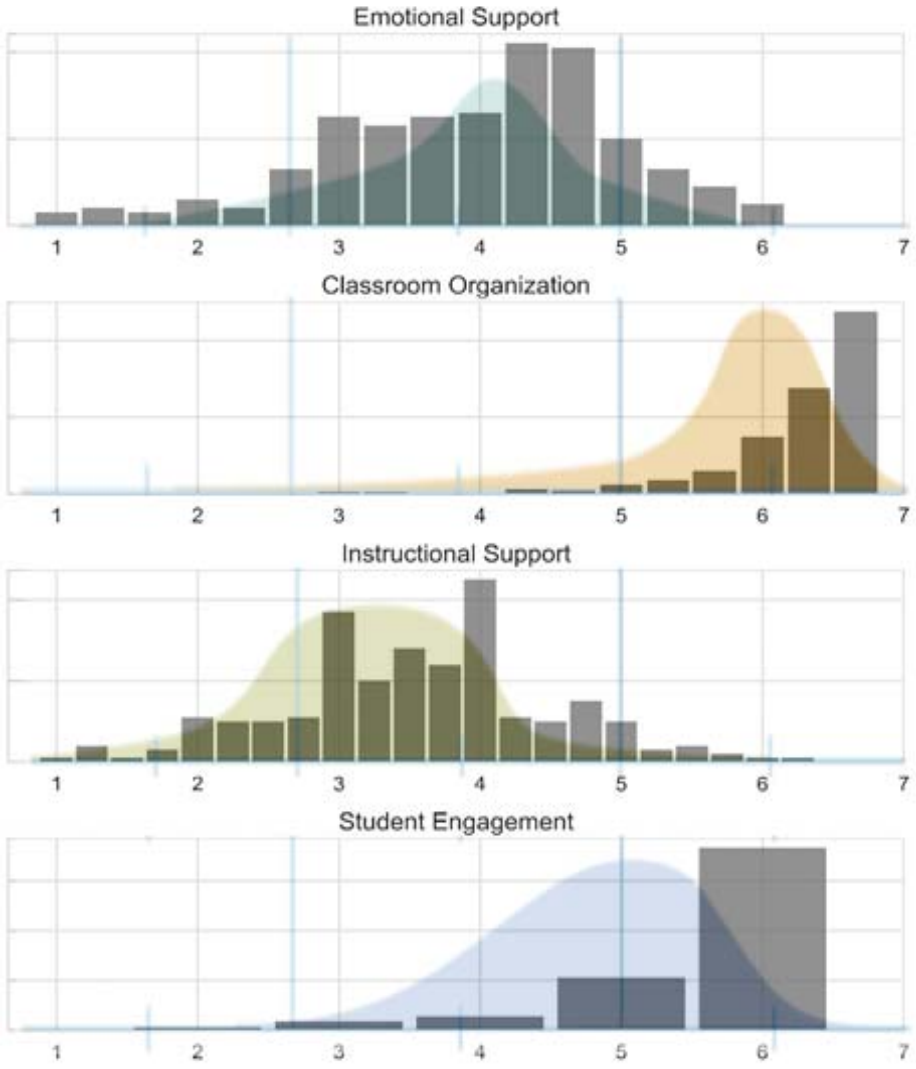
Figure 3: CLASS Scores across Domains, Guangdong Pilot Counties and Comparators



Source: Guangdong class observations; authors' calculations; Teachstone.

Figure 4 is a histogram of the domain scores overlaid with a smoothed result of other Upper Elementary CLASS results. The overall distributions of scores are very similar between the three pilot counties and data obtained from the MET study. The teachers observed in the three pilot counties had higher average scores than the comparator studies; the Classroom Organization and Student Engagement panels show higher distributions for the schools observed in this study.

Figure 4: Histogram of Guangdong Pilot Results Superimposed with Existing Results from MET study



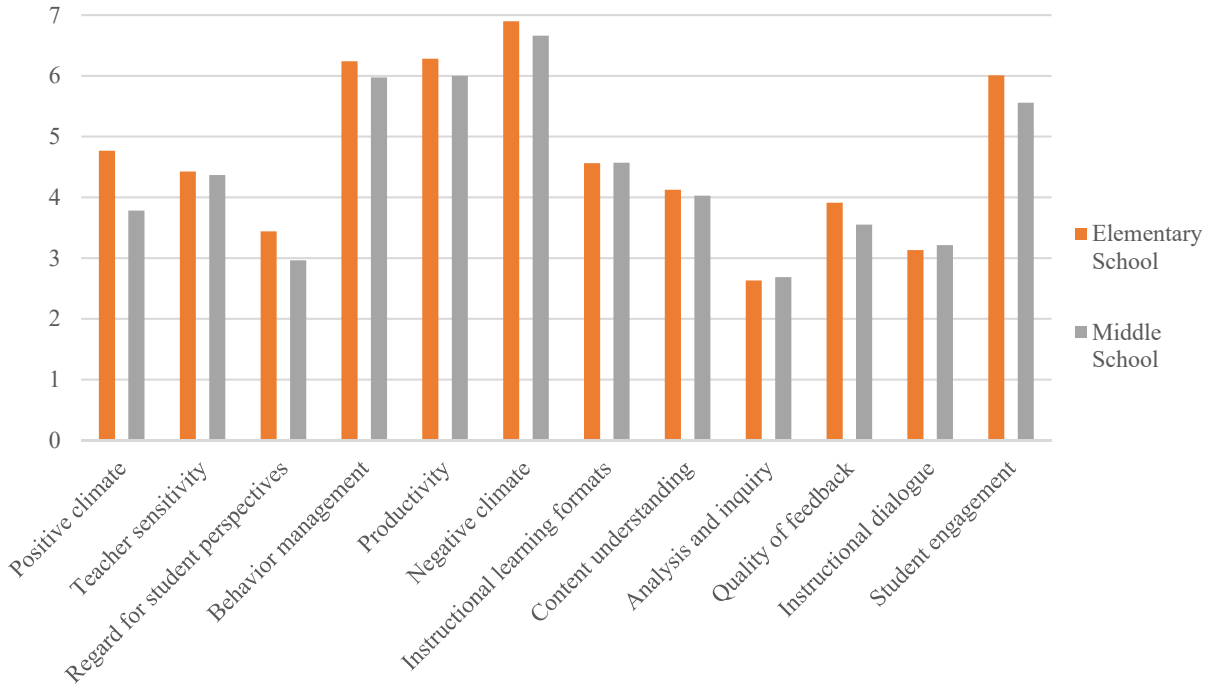
Source: Guangdong class observations; authors' calculations; Teachstone.

Note: Black bars are three pilot counties' values; the colored, smoothed overlays are the averages from applications of the Upper Elementary CLASS instrument under the MET study as provided by Teachstone.

3.1.2 Across dimensions

A detailed breakdown of the CLASS scores now examines the dimensions that make up the domains and reveals further disparities and similarities both within the Guangdong results and between the four studies (Guangdong and U.S.). Figure 5, displaying results from Guangdong, shows that teachers scored lowest on Regard for Student Perspectives, and on Analysis and Inquiry.

Figure 5: Analysis of Dimension Scores in Guangdong Pilot Counties



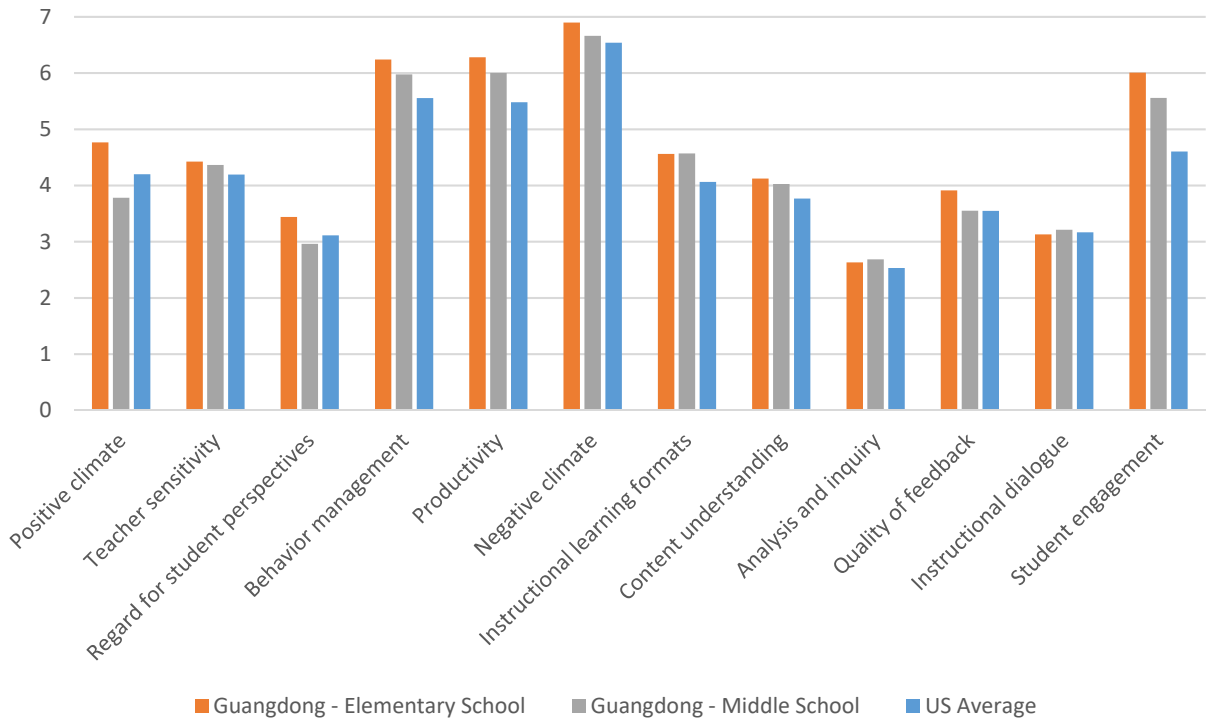
Source: Guangdong class observations; authors' calculations.

Figure 6 shows the dimension-level breakdown, as well as the U.S. comparator, which is an average of the results from the three United States–based studies. The dimensions weakest for the Guangdong classrooms are also the areas where U.S. classrooms are the weakest: Regard for Student Perspectives, and Analysis and Inquiry; a lack of Negative Climate is a strength for all classrooms.

When comparing these results, one should consider if raters' relative beliefs influenced their ratings at all. That is to say, whether a 4 in Content Understanding in Guangdong is equivalent in quantity and quality to a 4 in a U.S. classroom. If the 4 is relative to a pre-existing cultural baseline, it is possible that the intercultural comparisons may be harder. That does not, however, limit the utility of the tool for improving teacher training, as even a cursory glance at the data shows that the observed classrooms rank significantly lower in key areas. Further studies may benefit from a side-by-side comparison of videotaped classrooms from the United States and China to measure whether there are any issues on different baselines.

Finally, Student Engagement jumps out as an area in which the Guangdong classrooms are significantly more successful than U.S. classrooms (although this dimension is not included in any of the domain scores as it is its own category).

Figure 6: Average CLASS Scores for all Dimensions, Guangdong Pilot Counties and U.S. Comparator



Source: Guangdong class observations; authors' calculations; Teachstone.

Note: For comparability with the U.S. studies, the Negative Climate score has been reversed in the figure to make higher, better.

3.2 By county

This section examines the Guangdong pilot results by county. The three pilot districts, while all a part of the 16 project counties, varied widely by several measures, including economic prosperity and academic achievement. Were the differences in outcomes reflected in the CLASS scores, or was there relative equality in teaching quality? The results may surprise.

Table 7 provides a breakdown of the three pilot counties. Each county has data on total population, GDP per capita, and Zhongkao scores, both the absolute values and the rank out of the 16 project counties. All values are from 2014. The Zhongkao is a countywide test taken by all students at the end of 9th grade and serves as the best available comparator for cross-county educational achievement. These tests are not identical between counties, however, making it hard to compare directly. For this ranking, the percentage score was used to make comparisons between the project counties.

Table 7: Breakdown of Population, GDP per Capita, and Zhongkao Scores by County, 2014

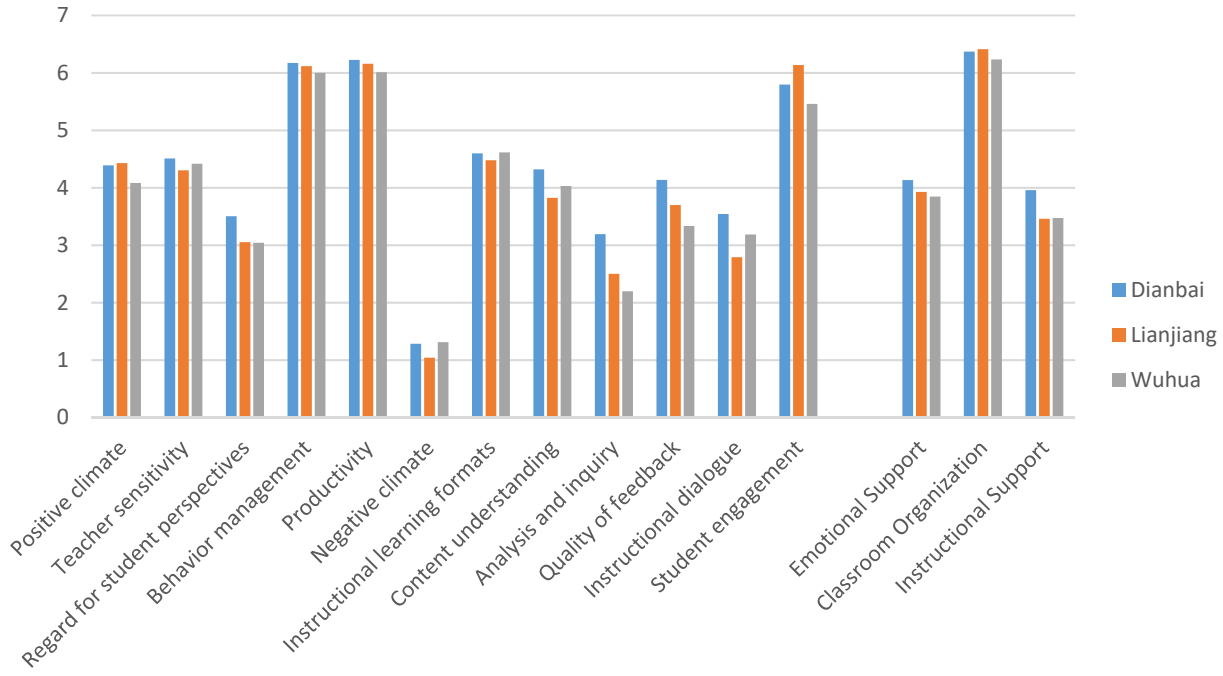
	Population		GDP per Capita (Y)		Zhongkao Score	
	Absolute	Rank	Absolute	Rank	Absolute	Rank
Dianbai	1,460,000	5	28,969	3	398/780	7
Lianjiang	1,485,000	3	24,006	8	409/920	16
Wuhua	1,363,000	7	11,452	16	426/800	3

Source: Data provided by Guangdong pilot counties, 2015.

While all counties have roughly equal populations, they differ greatly in economic output and academic achievement. Wuhua, the poorest of the 16 project counties, is ranked an impressive third among the counties by average Zhongkao score. Lianjiang, with triple the GDP per capita of Wuhua, is ranked last of the project counties by this score. Dianbai, third richest among the project counties, is right in the middle of the group for this score. For the 16 project counties, the overall correlation between GDP per capita and Zhongkao score (expressed as a percentage) is only 0.15. This is worth further exploration, and raises questions over the best way to rank the counties when evaluating the data.

Despite these differences in GDP per capita and educational attainment among the three counties, the average CLASS scores were all relatively similar (figure 7). Wuhua, which scored highest among the three counties in its Zhongkao results, scored the lowest in CLASS scores for Analysis and Inquiry and Quality of Feedback, as well as Student Engagement, Regard for Student Perspectives, and Positive Climate. It had the highest Negative Climate scores.

Figure 7: Average CLASS Scores by County



Source: Guangdong class observations; authors' calculations.

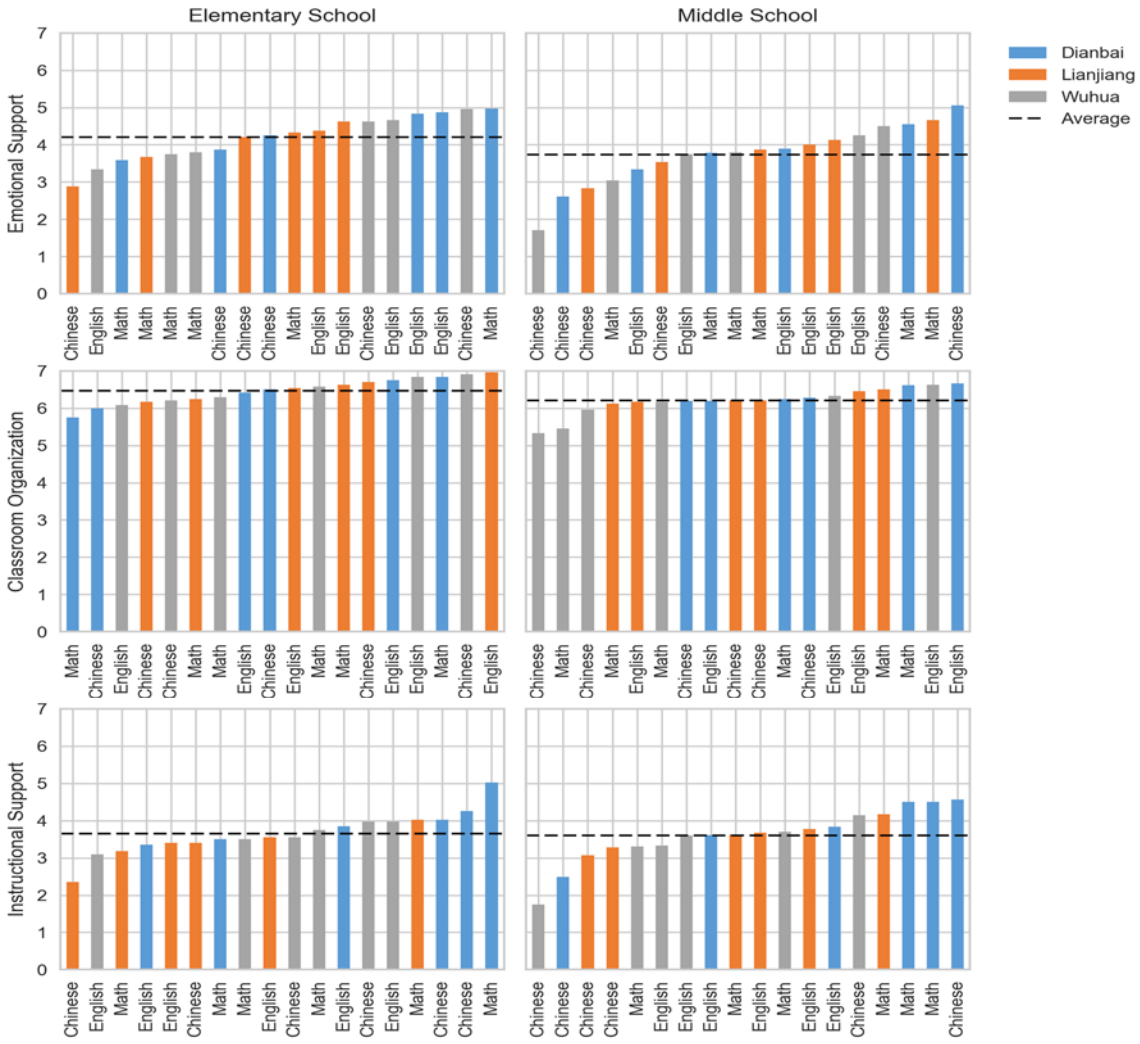
3.3 By teacher

Moving beyond county-wide averages, this section examines domain scores for all 36 teachers observed. In figure 8, the domain scores for the 18 elementary and 18 middle-school teachers are graphed, along with the average score for each domain. The color of the bar represents the county for each teacher, and the x-axis label represents the subject that the teacher taught. By plotting each teacher's average score across the domains, it is easier to understand the distribution of the scores by district.

The averages for the three domains are relatively close, but that belies a larger difference in the individual scores. For example, in middle school, the range of averages is 1.7 to 5.1—the maximum is three times the minimum. The range for Instructional Support is nearly as large—the maximum is 2.6 times the minimum. Elementary school had a smaller range of scores among the three domains, with Instructional Support the largest—the strongest teacher scores 2.1 times the lowest.

Even though the averages for Instructional Support in elementary and middle school are both nearly 4, less than a third of teachers are at or above a 4. This leads to several practical questions: Do officials aim for a higher level of competency, knowing that many teachers may not reach that level? How do they leverage the experience of teachers who are already above that level?

Figure 8: Teachers' Average CLASS Scores by Domain and County



Source: Guangdong class observations; authors' calculations.

3.4 The relationship between teacher beliefs and teaching quality

Teachers were given a 58-question survey on their beliefs related to teaching and learning. It used Likert-scale questions to gauge attitudes toward their work, their role in the classroom, and student achievement.

3.4.1 Teacher beliefs

The systematic study of teachers' conceptions of teaching interests researchers because many consider them to play a significant role in shaping teachers' characteristic patterns of instructional behavior (Thompson 1992, pp. 130–131). Wilson et al. (2002) noted what appears to be a basic premise of mainstream research into teachers' beliefs: "Over the last 15 years, there has been a considerable amount of research on teachers' beliefs based on the assumption that what teachers believe is a significant determiner of what gets taught, how it gets taught, and what gets learned in the classroom" (p. 128). Beliefs act as a filter through which teachers make their decisions rather than just relying on their pedagogical knowledge or curriculum guidelines (Clark and Peterson 1986). Teaching practices shape what happens in the classroom, but beliefs can help explain why they are used.

Beliefs are important to understand from a policy and teachers' professional development perspective. Cooney et al. (1998) state that "an analysis of belief structures, attention to the intensity with which beliefs are held, and the nature of the evidence that supports beliefs can provide a forum by which our teacher education programs will be better able to address issues of reform" (p. 331). Thompson (1984) focused on the negative consequence of not taking teacher mathematical beliefs into account when stating that "A failure to recognize the role that the teachers' conceptions might play in shaping their behaviour is likely to result in misguided efforts to improve the quality of instruction in the schools" (Thompson 1984, p. 106). Kagan (1992) argues that beliefs may be the clearest measure of a teacher's professional growth.

3.4.2 Teacher orientations

A teacher's system of beliefs may be considered to form what have been termed orientations. Askew et al. (1997) defined three teacher orientations of *transmissionist*, *connectionist*, and *discovery*. Renne (1992) developed a Purpose of Schooling/Knowledge matrix to conceptualize teachers' conceptions of teaching and learning mathematics, which distinguished between *school knowledge–oriented* (believing that teaching is an act of passing information on to others while learning involves the process of reproducing that information) vs. *child development–oriented* (more likely to consider children's needs and characteristics as the primary factors in instructional decision making).

The relationship and relative dominance or conviction among a teacher's beliefs should be recognized and considered. Thompson (1992) identifies three dimensions that relate to the ways beliefs are organized: their quasi-logical structure (primary vs. derivative), the degree of conviction with which they are held (central vs. peripheral), and their connectedness (clustered vs. isolated).

3.4.3 Teacher archetypes

The survey questions were grouped into three archetypes of teacher views and correlated against the classroom observations.

1. **Transmission**. These teachers believe that the role of the teacher is to act as a transmitter of knowledge, and students are supposed to receive that knowledge passively.
2. **Connection.** Teachers who score highly in terms of connection believe in highlighting relationships within the subject they are teaching, helping students to see the relationship between different lessons. It is a more student-centered approach.
3. **Discovery.** This is the most student-centered approach. Teachers provide a framework for knowledge and students are left to discover the lessons themselves, with the teacher's guidance and support.

To measure each teacher's support for a given archetype, the questions were coded within the teacher beliefs survey for their relation to that archetype. Depending on their degree of agreement or disagreement with a given belief, teachers were assigned points for a certain archetype. In this way, each teacher received a score indicating their implied adherence to each archetype. These scores were then correlated with the CLASS observation results, as well as with age, teaching experience, and number of students (table 8).

Table 8: Correlations between Teacher Archetypes and CLASS Results

	Transmission	Connection	Discovery
Age	0.09	0.16	-0.07
Teaching Experience	0.11	0.25	-0.03
No. of Students	-0.06	-0.11	0.06
Positive Climate	-0.34	0.17	0.40
Teacher Sensitivity	-0.10	0.04	0.12
Regard for Student Perspectives	-0.20	0.20	0.21
Behavior Management	-0.15	0.10	0.28
Productivity	-0.12	0.14	0.10
Negative Climate	0.00	-0.29	-0.10
Instructional Learning Formats	0.00	0.07	0.06
Content Understanding	0.16	0.15	0.16
Analysis and Inquiry	0.23	0.25	0.16
Quality of Feedback	-0.07	0.17	0.33
Instructional Dialogue	0.02	0.30	0.25
Student Engagement	-0.20	0.12	0.14
Emotional Support	-0.26	0.16	0.29
Classroom Organization	-0.11	0.18	0.19
Instructional Support	0.08	0.23	0.23

Source: Guangdong teachers' belief surveys; authors' calculations.

The more student-centered beliefs (connection and discovery) are more highly correlated with the student-centered actions, such as Positive Climate and Regard for Student Perspectives. More important, these results are also correlated with higher Instructional Support and Classroom Organization. Indeed, teachers who believe that their role in the classroom is to transmit information to students had a negative correlation with Emotional Support and Classroom Organization. This is important, as it is sometimes stated that while more teacher-centered beliefs may not support a positive classroom climate, they are more effective at ensuring that content is delivered effectively. High correlations between student-centered discovery and connection beliefs and Instructional Support undermines this claim, implying that a student-centered approach is also one that allows for improved Instructional Dialogue and Student Engagement.

4 Moving Forward

4.1 Evaluating results

While the Emotional Support and Instructional Support domains have a good deal of variation, the Classroom Organization and Student Engagement domains have very high scores across the board, with very little variation. With Classroom Organization, 86 percent of all coded sessions received a 6 or higher

and less than 2 percent received a score of less than 5. Similarly, with Student Engagement, 72 percent received a 6 or higher and less than 10 percent received a score of less than 5. With such little variation, there would be very little possibility of capturing measurable changes in these domains.

More variation appears when examining by specific dimensions. While a few showed little variation (e.g. Negative Climate, Productivity, and Behavior Management), others had a good deal of variation with even distributions (e.g. Quality of Feedback, Emotional Support, and Teacher Sensitivity). There were also scores that were particularly low and where measurable improvements could emerge through training activities (e.g. Instructional Dialogue, Analysis and Inquiry, and Regard for Student Perspectives).

The most interesting comparisons are weaker domains and dimensions, and the international comparisons. Observed teachers received higher scores in Classroom Organization, and lower scores in Emotional Support and Instructional Support. This is comparable with other CLASS assessments in the United States. Elementary-school teachers scored marginally higher than middle-school teachers in areas of Emotional Support, while teachers with more experience tended to have higher scores in Instructional Support.

On interregional comparisons, the three pilot counties did not exhibit significant differences in scores. This may be because the differences between the counties were not large enough—a comparison with teachers in the best schools in Guangzhou or Shanghai might reveal larger differences. A similar result occurred on the urban/rural distinction: given the overall level of development of the pilot counties, the distinctions between the two types of schools were not significant in affecting scores.

There were minor differences based on teacher type, although it was not as clear cut as saying that backbone teachers always outperformed new teachers on all dimensions.

The pilot study showed that the CLASS tool could be used in the context of the three pilot counties. The exploration of differences between teacher types, as well as by county and training, did not reveal as much difference as originally posited. To further explore these differences, a larger sample size would be necessary.

4.2 Contextualizing results through participant feedback

In June 2017, a workshop was held to discuss these findings and gather input from participants on their experiences of using CLASS and their perception of the tool's applicability. The participants included coders, teachers who were observed, Department of Education officials, and university personnel. There was strong agreement on the usefulness and general applicability of CLASS, but issues were raised on the challenge of coding specific dimensions. Results of the coding indicated that there were greater differences in scoring among pairs of coders for Teacher Sensitivity, Regard for Student Perspectives, Instructional Dialogue, and Analysis and Inquiry. Participants indicated that such discrepancies could in part be reduced through more intensive training specifically on these dimensions to develop a better understanding of the indicators. But some participants indicated that cultural aspects were an important factor in coding discrepancies for Teacher Sensitivity and Regard for Student Perspectives. They suggested that these dimensions should be revisited to consider how the instructions for coding might be modified to be more applicable to the Guangdong context. Overall, the participants in the workshop felt that the pilot study provided useful information on teacher strengths and areas for improvement.

They also felt that the lessons learned from the pilot study could also be used to improve the existing system for delivering teacher training. While continuing to develop and deliver teacher training as it has always done, the government could use the results of classroom observation as a results measure under a results-based financing mechanism. Doing so would give the government a consistent yardstick with which to provide content developers and service providers (often the same university) an added incentive of receiving a bonus based on the effectiveness of their training programs. Effectiveness could be gauged by improvements in teacher knowledge (as measured by pre- and post-training assessments) and by changes

in teaching practices (as measured by pre- and post-training classroom observations). The design of such a mechanism is detailed in Appendix C, with a conceptual framework and key areas of consideration.

4.3 Recommendations

Two sets of recommendations emerge from this pilot study. The first focuses on policy and the second on future research.

How CLASS and supplemental tools can be incorporated into routine assessments of teaching quality

Participants at the June 2017 workshop made the following suggestions for subsequent use of the CLASS tool in routine assessments of teaching quality (particularly in the context of the proposed Guangdong Compulsory Education project):

- As a first step, existing teacher training content should be examined through the lens of the CLASS instrument to identify potential gaps between what the training covers and what the tool assesses to determine what modifications are needed to improve the training.
- Second, the effectiveness of revised teacher training should be assessed based on the subset of CLASS dimensions the revised training is meant to improve—rather than all domains. As a preliminary list, participants felt that the following dimensions are likely to be most relevant:
    - Student Engagement
    - Content Understanding
    - Instructional Dialogue
    - Instructional Learning Formats
    - Positive Climate
    - Teacher Sensitivity
    - Regard for Student Perspectives
    - Quality of Feedback
- Third, the indicators for the Teacher Sensitivity and Regard for Student Perspectives dimensions should be examined in more detail to consider the cultural applicability and educational system relevance to determine whether modifications should be made or whether the dimensions should be dropped.
- While participants felt it important to maintain the integrity of the CLASS instrument, they felt there was scope to include additional factors that are relevant to the Guangdong context, but not adequately captured under the existing dimensions, such as whether teachers are supporting character development among students and whether or not they are effectively utilizing existing ICT resources.

Areas for further data collection and research on the CLASS tool in China

Given the scale of this pilot study, it was not always possible to capture significant differences across counties or groups of teachers. Further data collection and research are therefore needed to confirm some of the patterns observed. Specifically:

- Further work must be done to determine whether the results obtained in Guangdong for the domain of Instructional Support are related to difficulty of measuring some of the domains (including Analysis and Inquiry and Instructional Dialogue), or to an overall weakness on the part of teachers in providing adequate, student-centered instructional support.
- Both U.S. and Guangdong classrooms score poorly on some dimensions. Often these are the same dimensions, but the relative levels are higher in Guangdong than in the United States. Further studies may benefit from a side-by-side comparison of videotaped classrooms from the two countries to better understand what might be driving these observed differences.

5 REFERENCES

Askew, M., Rhodes, V., Brown, M., Wiliam, D. and Johnson, D. (1997). Effective teachers of numeracy. *London, King's College London*.

Beswick, K. (2005). The beliefs/practice connection in broadly defined contexts. *Mathematics Education Research Journal*, 17 (2), 39-68.

Bruns, Barbara, Soledad De Gregario and Sandy Taut (2016). RISE Working Paper 16/009—Measures of effective teaching in developing countries, September, 2016

China's grim rural boarding schools (2017). *The Economist*. Accessed June 12, 2017. http://www.economist.com/news/china/21720603-millions-children-countryside-attend-wretched-schools-far-home-chinas-grim-rural

Clark, C. and Peterson, P. (1986). Teachers' thought processes' in MC Wittrock (ed.): Handbook of Research on Teaching.

Conn, K. (2014). *Identifying Effective Education Interventions in Sub-Saharan Africa: A meta-analysis of rigorous impact evaluations* (Doctoral dissertation, Columbia University).

Cooney, T.J., Shealy, B.E. and Arvold, B. (1998). Conceptualizing belief structures of preservice secondary mathematics teachers. *Journal for Research in Mathematics Education,* 29 (3), pp. 306-333.

Evans, D., and A. Popova. (2015). What really works to improve learning in developing countries? An analysis of divergent findings in systematic reviews. *World Bank Policy Research Working Paper*, (7203).

Fang, Z. (1996). A review of research on teacher beliefs and practices. *Educational Research,* 38 (1), pp. 47-65.

Guangdong Department of Education. (2015). *Education and Economic Indicators* [Data set].

Hamre, Bridget K., Robert C. Pianta, Margaret Burchinal, Samuel Field, Jennifer LoCasale-Crouch, Jason T. Downer, Carollee Howes, Karen LaParo, and Catherine Scott-Little (2012). "A course on effective teacher-child interactions: Effects on teacher beliefs, knowledge, and observed practice." *American Educational Research Journal* 49, no. 1 pp. 88-123.

Howe, K. R. (1988). Against the quantitative-qualitative incompatibility thesis, or, Dogmas die hard. *Educational Researcher*, 17, 10–16.

Hu, Bi Ying, Xitao Fan, Chuanhua Gu, and Ning Yang (2016). "Applicability of the Classroom Assessment Scoring System in Chinese preschools based on psychometric evidence." *Early Education and Development* 27, no. 5 pp. 714-734.

Kagan, D.M. (1992). Implication of research on teacher belief. *Educational psychologist,* 27 (1), pp. 65-90.

Kremer, M., C. Brannen, and R. Glennerster. (2013). The challenge of education and learning in the developing world. *Science*, *340*(6130), 297-300.

Li Keqiang (2016). *Report on the Work of the Government.* Delivered at the Fourth Session of the 12th National People's Congress of the People's Republic of China on March 5, 2016.

Lu, M., Loyalka, P., Shi, Y., Chang, F., Liu, C., & Rozelle, S. (2017). The Impact of Teacher Professional Development Programs on Student Achievement in Rural China.

Martínez-Rizo, Felipe (2012). Procedures for study of teaching practices. Literature review. RELIEVE, v. 18, n.1, art.1. http://www.uv.es/RELIEVE/v18n1/RELIEVEv18n1_1eng.htm

McEwan, P. J. (2015). Improving Learning in Primary Schools of Developing Countries A Meta-Analysis of Randomized Experiments. *Review of Educational Research*, *85*(3), 353-394.

Medley, Donald M. & Harold Mitzel (1963). Measuring classroom behavior by systematic observation. Gage, N.L. Ed. Handbook of Research on Teaching. Chicago. Rand McNally, pp. 247-328.

Ministry of Education, China. (2011). Ten Year Plan for the Development of Education Informatization (2011-2020) [教育部发布教育信息化十年发展规划 (2011－2020)]. http://goo.gl/IJBgIO

Pakarinen, Eija, Marja-Kristiina Lerkkanen, Anna-Maija Poikkeus, Noona Kiuru, Martti Siekkinen, Helena Rasku-Puttonen, and Jari-Erik Nurmi (2010). "A validation of the classroom assessment scoring system in Finnish kindergartens." *Early Education and Development* 21, no. 1 pp. 95-124.

Renne, C. (1992). Elementary school teachers' views of knowledge pertaining to mathematics. *Annual Meeting of the American Research Association, San Francisco, CA.*

Rosenshine, Barak & Norma Furst (1973). The use of direct observation to study teaching. En Travers, Robert M. W. Ed. Second Handbook of Research on Teaching, pp. 122-183. Chicago: Rand McNally College Publ. Co.

Stallings, J. A., & Mohlman, G. G. (1988). Classroom observation techniques. In J. P. Keeves (Ed.), Educational research, methodology, and measurement: An International handbook (pp. 469-474). Oxford, England: Pergamon.

Statistics Bureau of Guangdong Province. (2014). *Guangdong Statistical Yearbook* [Data set]

Stigler, J. W., Gallimore, R. & Hiebert, J. (2000). Using video surveys to compare classrooms and teaching across cultures: Examples and lessons from the TIMSS video studies. *Educational Psychologist*, 35(2), 87-100.

Swan, M., (2006). National research and development centre for adult literacy and numeracy and National institute of adult continuing education (England/Wales), *Collaborative learning in mathematics: a challenge to our beliefs and practices.* National Institute of Adult Continuing Education.

Thompson, A.G. (1984). The relationship of teachers' conceptions of mathematics and mathematics teaching to instructional practice. *Educational studies in mathematics,* 15 (2), pp. 105-127.

Thompson, A.G. (1985). Teachers' conceptions of mathematics and the teaching of problem solving. In E. A. Silver (Ed.), *Teaching and learning mathematical problem solving: Multiple research perspectives* (pp. 281-294). Hillsdale, NJ: Lawrence Erlbaum.

Thompson, A.G. (1992). Teachers' beliefs and conceptions: A synthesis of the research. *Handbook of research on mathematics teaching and learning,* 127, pp. 146.

Tsang, M.C., and Y. Ding. (2005). Resource utilization and disparities in compulsory education in China. *China Review*, 1-31.

von Suchodoletz, Antje, Anika Fäsche, Catherine Gunzenhauser, and Bridget K. Hamre (2014). "A typical morning in preschool: Observations of teacher–child interactions in German preschools." *Early Childhood Research Quarterly* 29, no. 4 pp. 509-519.

Wen J. Peng, Elizabeth McNess, Sally Thomas, Xiang Rong Wu, Chong Zhang, Jian Zhong Li, Hui Sheng Tian (2014). Emerging perceptions of teacher quality and teacher development in China, *International Journal of Educational Development*, Volume 34, Pages 77-89, http://dx.doi.org/10.1016/j.ijedudev.2013.04.005

Wilson, M., Cooney, T., Leder, G., Pehkonen, E. and Toerner, G. (2002). Beliefs: A hidden variable in mathematics education?

Zhang, D., X. Li, and J. Xue. (2015). Education Inequality between Rural and Urban Areas of the People's Republic of China, Migrants' Children Education, and Some Implications. *Asian Development Review*.

Zhang, X., and R. Kanbur. (2005). Spatial inequality in education and health care in China. *China Economic Review*, *16*(2), 189-204.

APPENDIX A EXTENDED RESULTS DISCUSSION

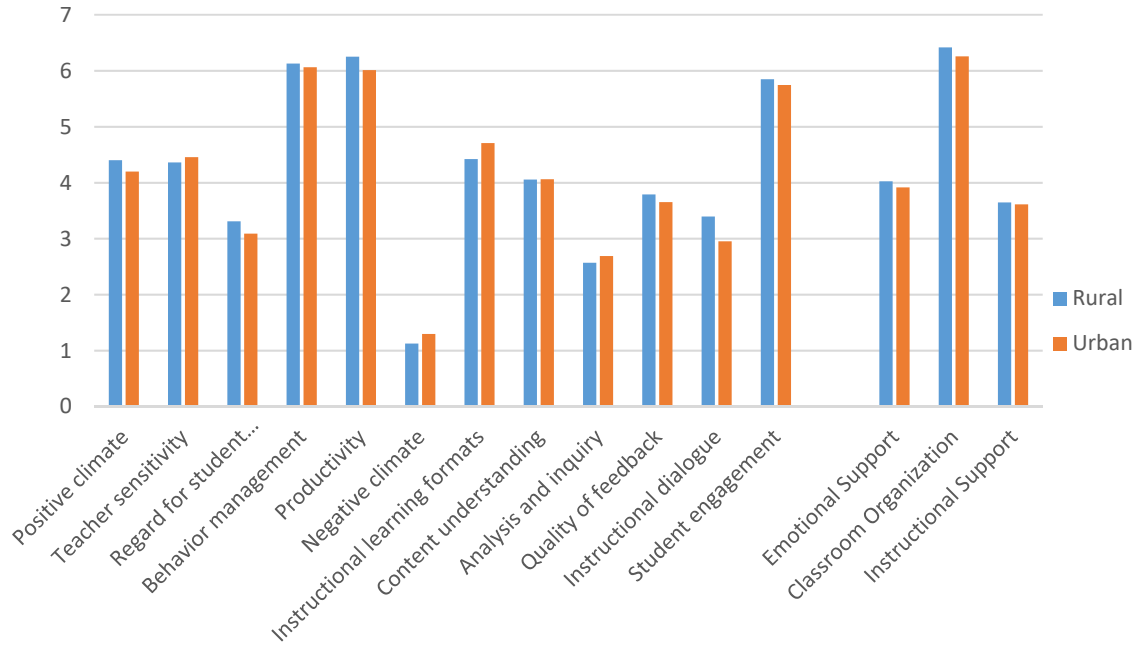
This appendix provides further avenues of analysis for the CLASS results, including differences by teacher experience and school location. This analysis did not provide much insight into the usability of CLASS in adjusting teacher training programs. Additionally, given the sample size for certain levels of analysis, these results should be seen only as a starting point for further testing and observation, rather than conclusive evidence of differences in teacher groups.

A.1 By school type

While all schools observed are in Guangdong's poorer counties, the conditions within each of these counties vary depending on whether one is in the countryside or in the county seat. For this study, half of the schools observed are in county seats, while half are in the countryside. Despite this distinction between urban and rural in our data set, it is likely that *all* schools observed would be classified as rural relative to true urban centers such as Guangzhou, Shanghai, or Beijing.

While the gap between urban and rural student achievement is clear, as noted in the introduction in relation to Programme for International Student Assessment (PISA) scores, as well as in Chinese college attendance and other indicators, the gap in teaching quality is not necessarily as simple. Often, teachers who graduate from teachers' colleges, known in China as "normal universities," are sent to rural classrooms for some years before being able to transfer to more urban schools. Likewise, high-performing vice principals and administrators from urban schools who wish to be promoted are often assigned to rural and under-resourced schools to share best practices. Finally, China has been implementing a program of rural school consolidation, merging several smaller schools into one larger school to leverage resources. All this is to say that there is often a large amount of support for rural schools, despite other challenges. Younger, recently graduated teachers benefit from the latest pedagogical instruction, and experienced principals can support a staff looking to improve an under-resourced school.

Figure A.1: Differences between Urban and Rural Classrooms

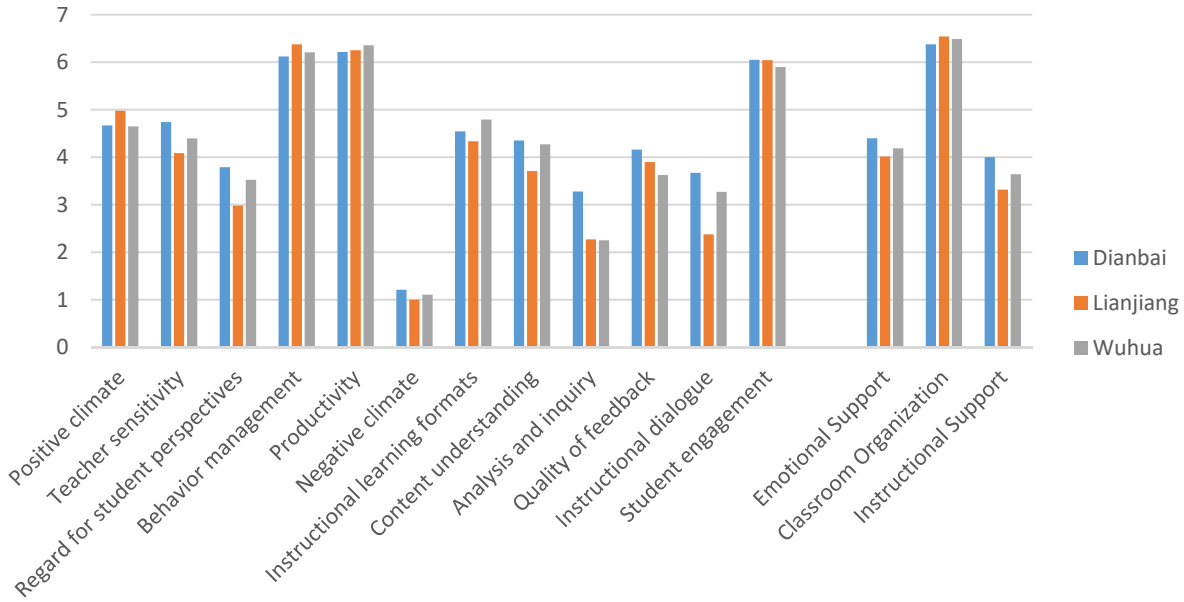


Source: Guangdong class observations; authors' calculations.

Figure A.1 shows that rural scores were slightly higher than urban classrooms across all three domains—Emotional Support, Classroom Organization, and Instructional Support—although only by a tenth of a point. With only 18 teachers and six schools for each category, it is difficult to draw any significant conclusions when the data are so similar. While the local counties are able to differentiate what urban and rural schools look like within their particular context, from the viewpoint of Beijing and Shanghai, all schools observed would likely be considered rural. With that in mind, it is understandable that there is not a significant difference in observed teacher behaviors.

A breakdown of scores by district for elementary and middle schools shows a similar trend in relative scores between the three counties.

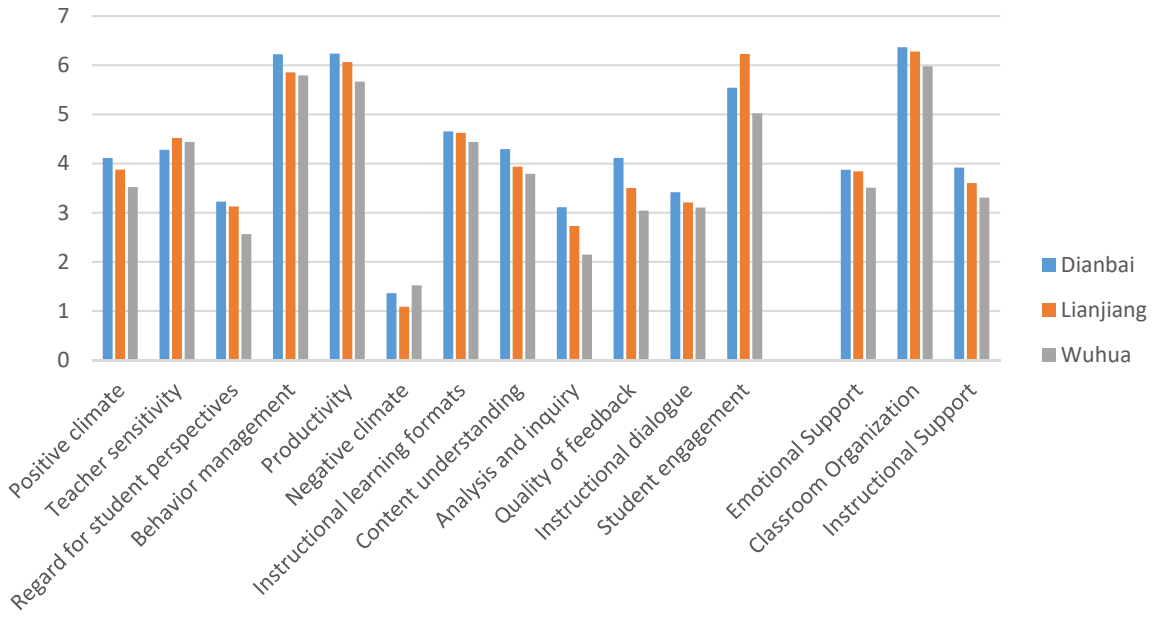
Figure A.2: Average Elementary School CLASS Scores by County



Source: Guangdong class observations; authors' calculations.

In figure A.2, Dianbai stands out, particularly in Instructional Support, where it outscores the other two counties in several dimensions.

Figure A.3: Average Middle-School CLASS Scores by County



Source: Guangdong class observations; authors' calculations.

Figure A3 shows Dianbai again standing out in almost all areas.

A.2 By teacher type (new, backbone candidate, backbone)

Moving to the teachers themselves, differences between three types of teachers were examined: new teachers, teachers who had been identified as backbone candidates, and backbone teachers. Backbone teachers are those who have been formally identified as leaders among their fellow teachers, often holding open classes and having demonstrated many years of professional distinction. For the purposes of this section, those teachers that have been identified as potential backbone teachers will be referred to as "candidates". Table A.1 summarizes the average ages and teaching experience of the three teacher types. There are 12 teachers for each type.

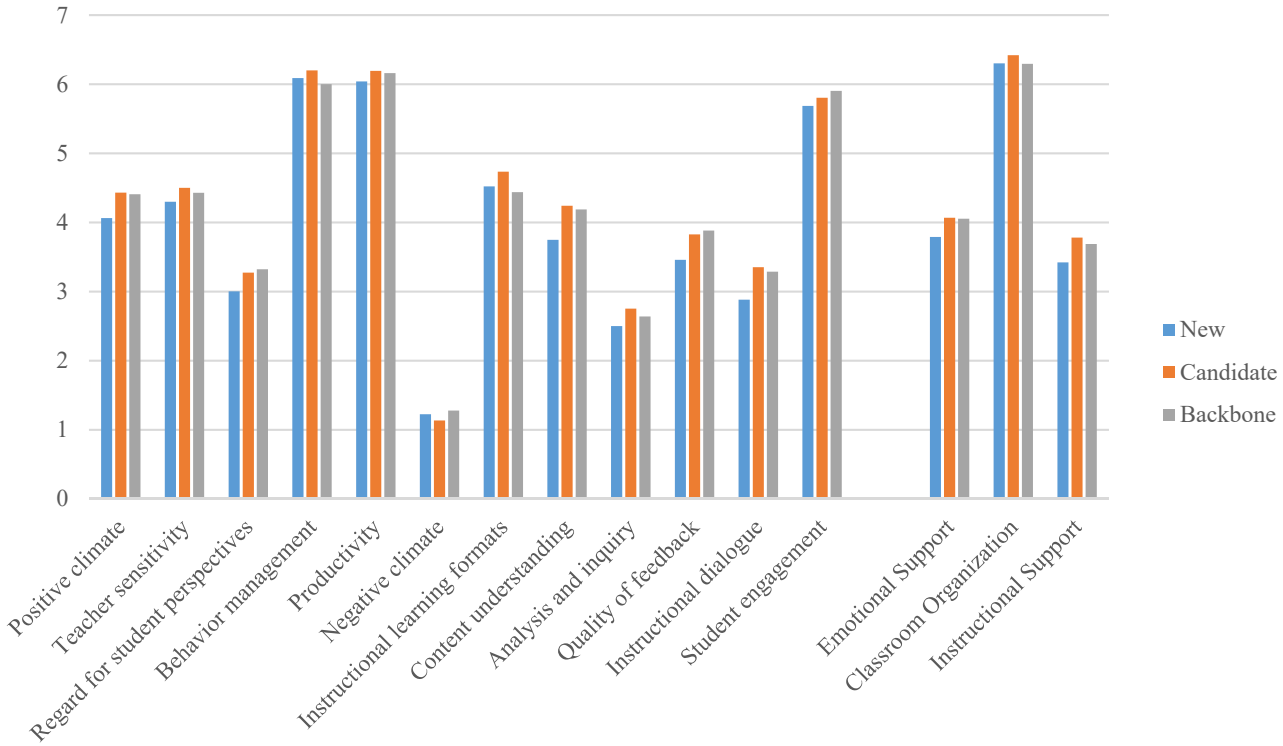
Table A.1: Average and Range of Age and Teaching Experience by Teacher Type

	Age (Range)	**Teaching Experience**
New	30.7 (26-38)	7.8 (1-19)
Candidate	36.7 (29-44)	15.5 (7-25)
Backbone	38.0 (28-47)	15.6 (4-25)

Table A.1 shows that on average, new teachers are the youngest and have the least experience, while backbone teachers are the oldest and most experienced. The ranges, however, show that the picture is somewhat more complicated, with significant overlap existing in ages and experience between the different categories of teachers.

Figure A.4 shows the scores for each type of teacher, along with the overall scores for all teachers. As seen in other breakdowns above, the differences in scores are not very large among the three different categories.

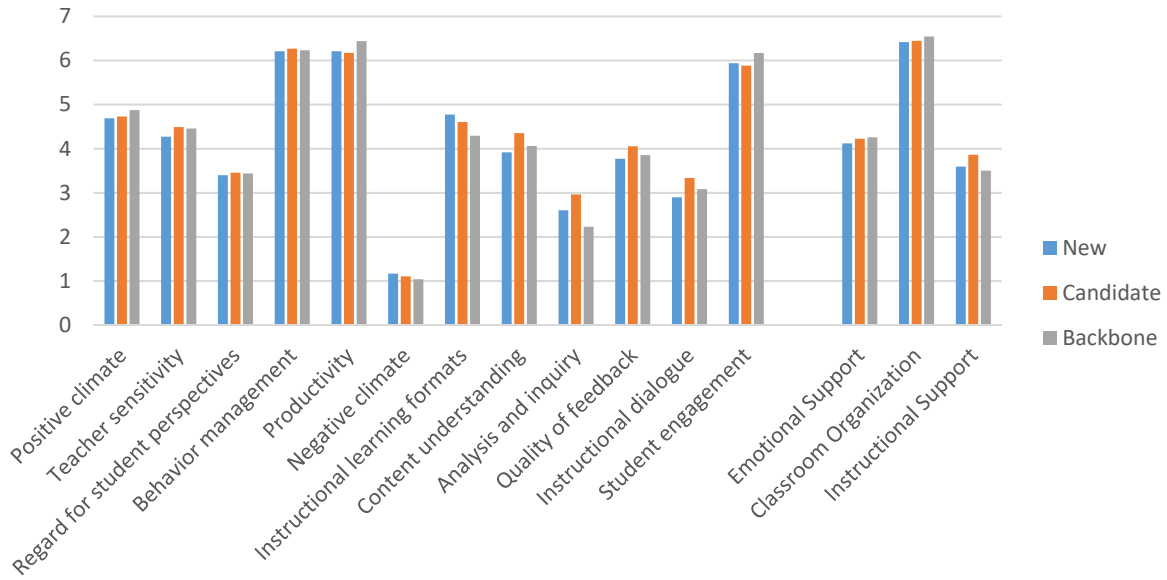
Figure A.4: Differences between Teacher Types



Source: Guangdong class observations; authors' calculations.

Candidate teachers stand out among the three categories, scoring the highest along all three domains. New teachers lag in Emotional Support and Instructional Support, while ranking equal to the other teachers in Classroom Organization. Backbone teachers have a slight edge in Student Engagement, Regard for Student Perspectives, and Quality of Feedback.

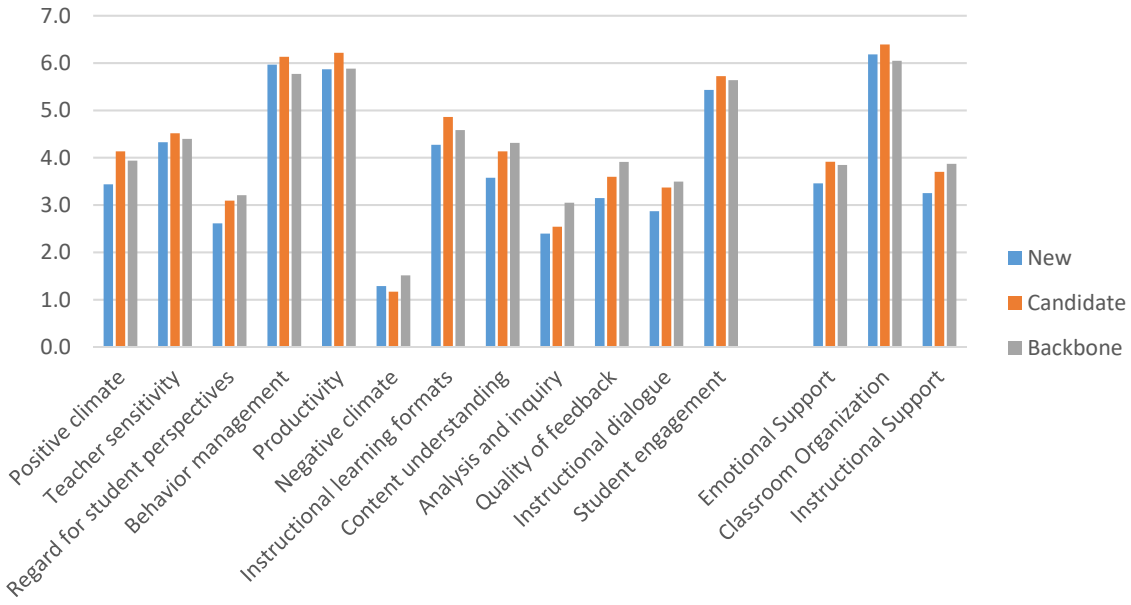
Figure A.5: Differences between Teacher Types in Elementary School



Source: Guangdong class observations; authors' calculations.

The differences between teacher types largely disappear in elementary-school classrooms, although there are some differences in Instructional Support dimensions (figure A.5). However, the sample above represents only 18 teachers, six for each teacher type. Additionally, instructional support is often the hardest domain to gauge effectively, making judgments regarding teacher types and instructional effectiveness difficult.

Figure A.6: Differences between Teacher Types in Middle School



Source: Guangdong class observations; authors' calculations.

A look at middle-school teacher experiences paints a different picture. As can be seen in figure A.6, new teachers in middle school have a more difficult time in Emotional Support and Instructional Support, scoring behind candidate and backbone teachers. Candidate teachers perform well in Emotional Support and Classroom Organization, while scoring slightly lower than backbone teachers in Instructional Support.

It appears as if the transition to elementary schools is significantly easier on new teachers than the transition to middle school. While new teachers are generally able to maintain Classroom Organization, they appear less able to effectively provide emotional or instructional support. There are of course many differences between the two contexts: class sizes are generally larger, there is more curricular pressure, and the content is more difficult for students, not to mention the difficulty of emotionally supporting teenagers.

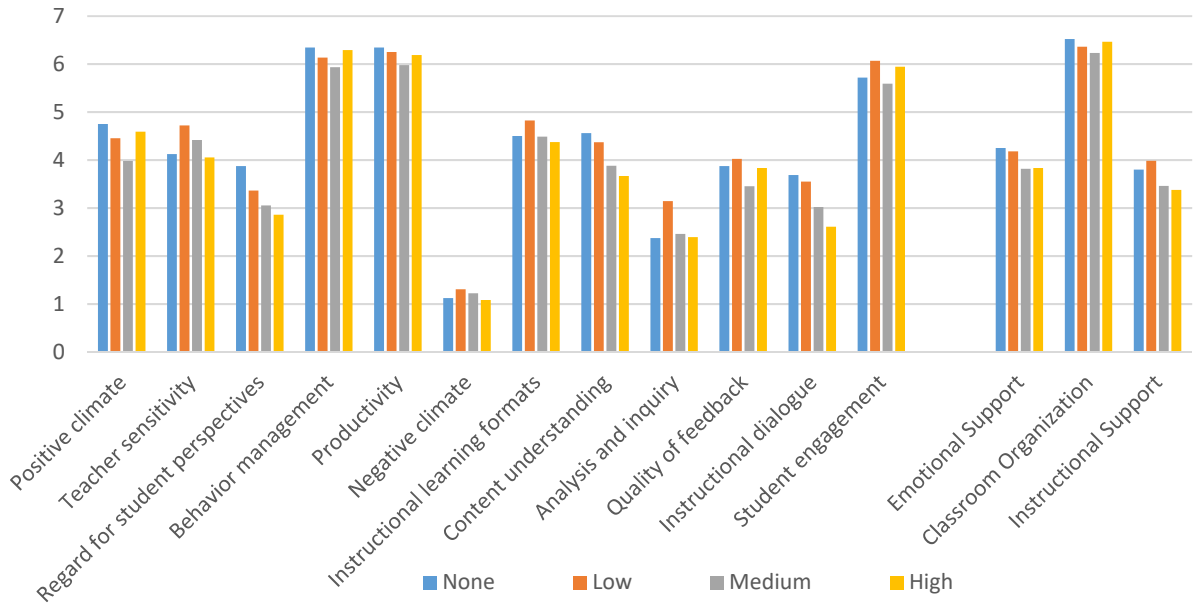
Within middle schools, the outperformance of candidate teachers stands out. Candidate teachers, supported by a more comprehensive training program as well as mentor teachers and administrators, may feel motivated in their professional development and increase their efforts in areas that may reflect favorably on their candidacy. Finally, there may be additional factors separate from those identified that play a more significant but yet unmeasured role. It is possible that the most impactful variable is related to a single subject, or a subset of rural candidate teachers versus urban candidate teachers. Given the number of teachers observed, it is not particularly useful to add additional filters to the data—one or two individual teachers would end up misleading rather than informing.

A.3 By training history

Along with the other information collected as part of the study, we also attempted to obtain data related to teachers' training history to measure potential effects of training on teaching practices. Flaws in the data collection methodology led to an incomplete sample being collected. We made two adjustments to deal

with the incomplete data. First, we grouped the teacher training data into four buckets: none, low, medium, and high, corresponding to the level of detail and completeness provided by the observed teachers. The results are in figure A.7.

Figure A.7: CLASS Scores Grouped by Reported Training

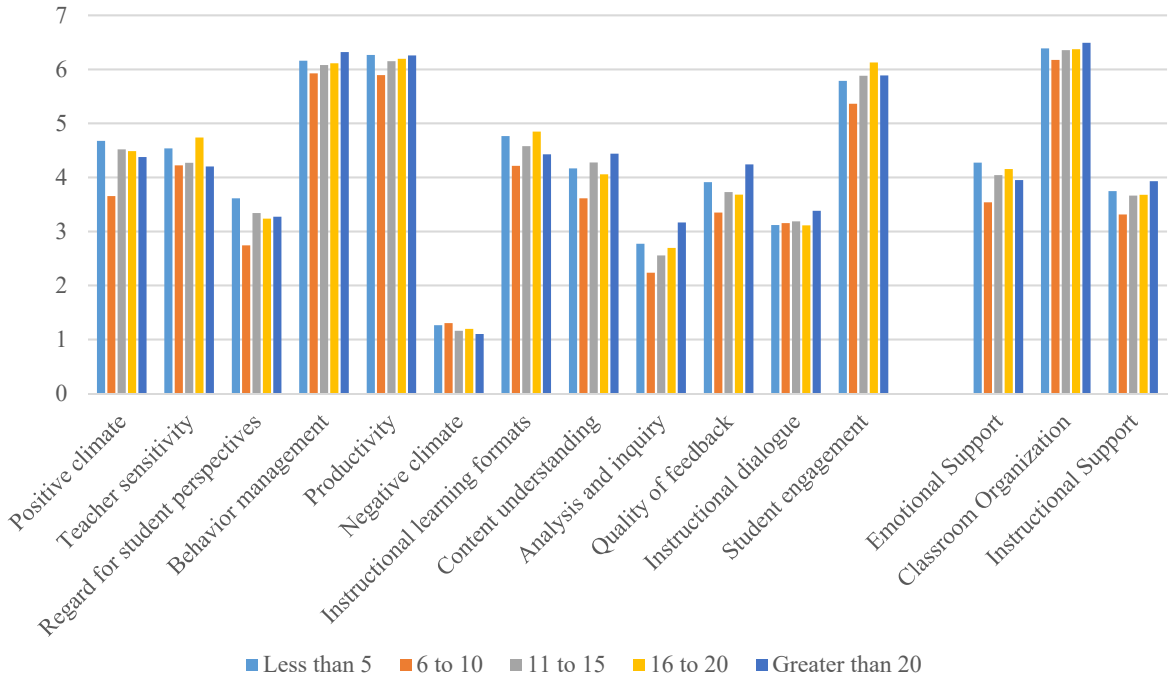


Source: Guangdong class observations; authors' calculations.

Overall, the grouping of teachers by reported training does not reveal any consistent patterns in CLASS scores. Indeed, the teachers who gave the most detail on the surveys regarding training history had the lowest average scores for Instructional Dialogue, Content Understanding, and Regard for Student Perspectives, as well as the lowest overall score for Instructional Support. Rather than interpreting this as a correlation between amount of training and teaching quality, it should be understood as a sign that more thorough responses on training history are needed for the next round of observations.

Second, given that teachers are generally required to attend annual training, and due to the wide range in experience by teacher type, we decided to group teachers by years of teaching experience rather than the self-reported training analysis done above. By grouping teachers for every five years of teaching experience (less than five years, six to 10 years, etc.), we created five similarly sized groups (figure A.8). This analysis straddles the line between training history and teacher type.

Figure A.8: CLASS Scores by Years of Teaching Experience



Source: Guangdong class observations; authors' calculations.

Note: Years of teaching experience were collected in the supplementary packets given to all observed teachers.

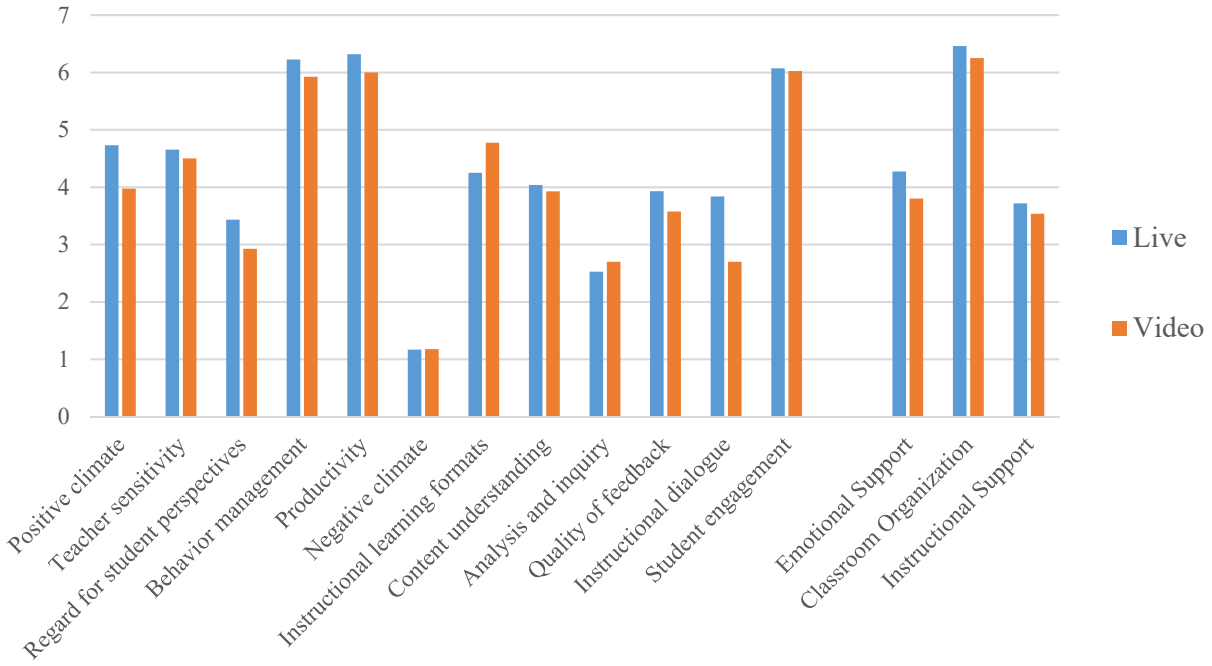
The least experienced teachers and the most experienced teachers stand out—the youngest teachers on Emotional Support and the oldest on Instructional Support. They are also the two strongest groups in Classroom Organization. The group of teachers with between six and 10 years of experience is weakest across all three domains. It is easier to hypothesize why the youngest, least experienced teachers may be the strongest at Emotional Support—newer pedagogical teachings emphasize improved teacher–student relations and fostering a positive classroom environment, while the most experienced teachers have decades of experience teaching content in an environment that stressed test results above all else. These are possible, but not definite, explanations for the trends in the data. It is more difficult to assess why there is a dip in the middle years.

A.4 By type of observation

Of the 36 teachers who were observed, 30 were videotaped while six were only evaluated in person. Five of the 30 videotaped were evaluated solely by videotape, while the remaining 25 were coded by a mix of live and video observations.

In this section, we compare the five that were coded only through video against the six that were observed only in person. In follow-up discussions, classroom raters noted that they felt that live observations allowed for better evaluation of the classroom environment as well as teacher-student interaction. The small sample size prevents us from drawing any significant conclusions, but it is notable that the live observations had higher scores across all dimensions except for Instructional Learning Formats and Analysis and Inquiry. Teachers observed live were scored nearly a point higher in Positive Climate and more than a point higher in Instructional Dialogue (figure A.9).

Figure A.9: Comparison between Live and Video Observations



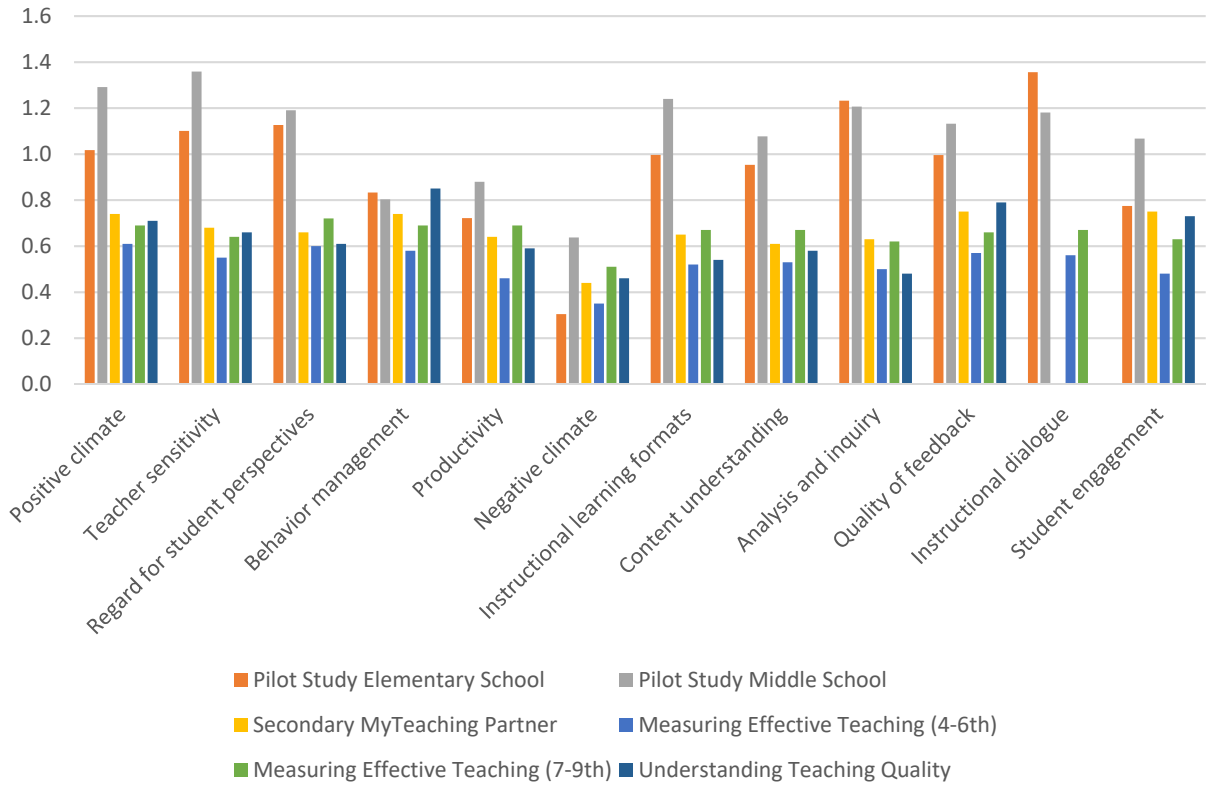
Source: Guangdong class observations; authors' calculations.

A.5 Standard Deviation of Scores

Finally, we calculated the standard deviation of the scores across all dimensions, and compared the results with those from the U.S. comparator studies (figure A.10).

The larger standard deviation could be introduced through two channels. First, it is possible that the raters were more willing to give a wider range of scores for a similar set of classrooms. Second, and more likely, is that the range of teachers observed in Guangdong was greater than in the U.S. studies. While on average these classrooms were quite similar in many regards, the range of quality among the Chinese teachers was greater. Given the greater differences in economic context and teacher training, this seems a possible explanation.

Figure A.10: Standard Deviation of CLASS Scores Across Dimensions



Source: Guangdong class observations; authors' calculations; Teachstone.

APPENDIX B STATISTICAL TESTS OF RELIABILITY

Three tests were used to gauge inter-rater reliability:

1. **Percent Agreement.** First, we measured the number of scores that were equal or adjacent (±1) between the two raters. For example, if one rater scored a certain dimension a 4, while their partner rated it a 3, 4, or 5, that would count as being in agreement. The results and analysis are presented below. This is in accordance with CLASS certification, which requires raters to score within one point of a master code to prove proficiency.
2. **Intraclass Correlation Coefficient (ICC).** Using a two-way mixed-agreement ICC we assessed the agreement between raters across subjects. Ratings range from -1 to 1, with higher ratings indicating higher levels of agreement between raters.
3. **Cohen's Kappa.** Linear and quadratic weighted kappas were used to measure agreement while taking into account the effects of chance. This statistical method evaluates the likelihood that agreement was due to observations of same behavior versus both raters randomly choosing same number. Ratings range from 0 to 1, with higher ratings indicating greater agreement not due to chance.

Three comparator studies are presented in addition to the three pilot counties to provide a reference for other implementations of the CLASS tool in upper elementary and junior secondary (table B.1).

Table B.1: Three Comparator Studies

Study	Number of Teachers	Grades	Raters	Supporting Organization	Subjects	Period
Measures of Effective Teaching (MET)	1,333	4–8	Graduate Students at University of Virginia (UVA)	Gates Foundation	Math and English	2009
Understanding Teaching Quality in Algebra Study (UTQ-A)	82	6–12	Trained Raters at Educational Testing Service	UVA, W.T. Grant and Spencer Foundations	Math and English	2009–11
Secondary MyTeachingPartner Study (S-MTP)	78	7–12	Graduate Students at UVA	UVA	No subject specified	2007–09

Percent Agreement

The simplest measure of agreement provides a quick way to gauge where raters had an easier or more difficult time in agreeing on observed behaviors. It is also in line with CLASS protocols, as the same measure is used to evaluate whether a rater has passed CLASS training.

Table B.2: Percent Agreement, Exact + Adjacent (%)

	Guangdong	S-MTP	MET	UTQ-A
Positive Climate	78.4	79.3	74.6	69.3
Teacher Sensitivity	58.4	73.8	72.6	65.0
Regard for Student Perspectives	64.8	78.6	67.9	70.2
Behavior Management	80.4	89.2	85.6	92.8
Productivity	91.5	84.3	82.4	87.2
Negative Climate	95.6	95.1	95.1	97.6
Instructional Learning Formats	72.0	80.3	78.2	73.9
Content Understanding	70.7	73.4	75.5	76.6
Analysis and Inquiry	66.4	73.7	71.5	84.3
Quality of Feedback	78.1	72.5	71.9	64.1
Instructional Dialogue	64.8	—	74.8	—
Student Engagement	83.1	81.2	78.4	82.0

Source: Guangdong observations; Teachstone.

Note: — indicates that the dimension was not scored for that study.

Table B.2 shows the average percent agreement between raters across the 12 CLASS dimensions. The raters in Guangdong had similar measures of agreement as those in three other studies that took place in a U.S. context. Certain dimensions proved harder for the Chinese raters to assess similarly, such as Regard for Student Perspectives, Teacher Sensitivity, and Analysis and Inquiry, perhaps due to differing definitions of what comprises effective Teacher Sensitivity in Chinese classrooms. On other measurements, such as Productivity and Student Engagement, the Chinese raters were in agreement a higher percentage of the time than in the U.S. context of the other three studies.

For future observations, trainers might be better to focus on more examples of those areas where raters had lower levels of agreement. While raters seemed to have similar ideas of what comprised a Positive Climate or a productive classroom, they were less sure as to Teacher Sensitivity or Regard for Student Perspectives.

Intraclass Correlation Coefficient

Table B.3: Intraclass Correlation Coefficient

	Guangdong	Guangdong*	S-MTP	MET
Positive Climate	0.58	0.60	0.43	0.42
Teacher Sensitivity	0.28	0.30	0.35	0.33
Regard for Student Perspectives	0.39	0.43	0.39	0.38
Behavior Management	-0.01	0.05	0.44	0.41
Productivity	0.30	0.34	0.38	0.28
Negative Climate	0.06	0.07	0.51	0.49
Instructional Learning Formats	0.38	0.38	0.39	0.35
Content Understanding	0.41	0.47	0.38	0.31
Analysis and Inquiry	0.33	0.34	0.33	0.26
Quality of Feedback	0.38	0.46	0.42	0.37
Instructional Dialogue	0.36	0.42		0.37
Student Engagement	0.36	0.39	0.41	0.27
Domains				
Emotional Support	0.67	0.70		
Classroom Organization	0.23	0.41		
Instructional Support	0.64	0.70		

Source: Authors' calculations; Teachstone.

Note: The Guangdong* column removes the two rater pairs least in agreement.

The two-way mixed-effect model for absolute agreement is a more stringent measure of inter-rater reliability, measuring not only differences in absolute scores, but also in ranking. Scores for the Guangdong assessment range from 0.28 to 0.58 across the dimensions, while domain scores range from 0.23 to 0.64 (table B.3). These ratings are in line with comparable studies, although none of the values across all studies would be considered as indicating extremely high consistency or agreement between raters per the most widely accepted standards of ICC tests.

Additionally, these ratings are averages for all rater pairs. Much as there were weaknesses in the percent agreement, there were also weaker pairs pulling down the averages. For example, if the two weakest pairs were removed from the Classroom Organization ICC score, the average would rise to 0.41, almost doubling the observed ICC.

Other CLASS studies indicate that more experience with the CLASS instrument may improve inter-rater reliability. In a paper on the applicability of CLASS in Chinese preschools, Hu et al. (2016) achieved average ICC scores of 0.82 to 0.92, indicating very high levels of consistency between raters. The authors, however, have been using CLASS for many years and have had more opportunity to discuss the measures.

Finally, given the range of subjects (math, English, Chinese), the difference in grades (4^{th} and 8^{th}), settings (urban and rural), and rating methods (in-person, videotaped), it is worth further investigation to see if any of these factors play a role in consistency and agreement between raters.

Table B.4: Cohen's Kappa

	Guangdong		S-MTP		MET		UTQ-A	
	Linear	Weighted	Linear	Weighted	Linear	Weighted	Linear	Weighted
Positive Climate	0.19	0.52	0.31	0.44	0.28	0.38	0.03	0.08
Negative Climate	0.04	0.06	0.41	0.55	0.36	0.47	0.07	0.13
Teacher Sensitivity	0.11	0.20	0.29	0.38	0.24	0.33	0.06	0.09
Regard for Student Perspectives	0.04	0.28	0.24	0.36	0.21	0.31	0.11	0.21
Behavior Management	0.10	0.03	0.38	0.51	0.31	0.48	0.08	0.14
Productivity	0.17	0.23	0.33	0.42	0.21	0.28	0.04	0.09
Instructional Learning Formats	0.06	0.29	0.3	0.39	0.21	0.33	0.02	0.09
Content Understanding	0.12	0.27	0.29	0.37	0.2	0.3	0.15	0.25
Analysis and Inquiry	0.06	0.23	0.22	0.31	0.13	0.2	0.08	0.17
Quality of Feedback	0.11	0.35	0.28	0.38	0.24	0.35	0.06	0.1
Instructional Dialogue	0.07	0.30	N/A	N/A	0.23	0.36	N/A	N/A
Student Engagement	0.14	0.29	0.33	0.39	0.19	0.29	0.07	0.14

Cohen's kappa is a measure of inter-rater reliability that compares the observed outcome against a statistically expected outcome. Like a correlation coefficient, it ranges from -1 to 1, with a value of 1 implying perfect agreement. Negative values imply that the raters agree less than would be expected by chance; that is, they are in disagreement. Different levels of interpretation exist for ranges of values; less than 0.2 can be considered poor; 0.2 to 0.4 fair; 0.4 to 0.6 moderate; 0.6 to 0.8 good; and 0.8 to 1.0 very good. Weighting Cohen's kappa considers the distance of disagreement between raters. Given that CLASS is more complicated than a yes/no coding, this has a significant effect on the calculation of the kappa.

The results from Guangdong did not imply strong agreement, but were in similar range to the comparator studies, and quite higher than many results from the UTQ-A study (table B.4). Guangdong's weighted Cohen's kappa ranged from 0.03 to 0.52, indicating low to moderate agreement distributed among the dimensions.[6] Given the subjective nature of classroom observations, as well as larger range of possible outcomes, it is not surprising that the Cohen's kappa for CLASS is lower than might be observed in a simple yes/no diagnosis.

Across the three measures of inter-rater reliability, the analysis reveals sufficient reliability to proceed with further analyses.

CLASS Reliability

Once the reliability of the raters was established, we then tested the internal consistency of the CLASS measures. Key tests included:

1. Correlations among CLASS dimensions and domains.
2. Cronbach's alpha of the three CLASS domains.

[6] Subsequent discussions with the coders revealed uncertainty regarding the coding of some dimensions; adjustments to the training program for coders should help improve the scores in the future.

Correlations among CLASS dimensions and domains

Table B.5: Correlations between CLASS Dimensions

	1	2	3	4	5	6	7	8	9	10	11
1. Positive Climate											
2. Teacher Sensitivity	0.60										
3. Regard for Student Perspectives	0.72	0.59									
4. Behavior Management	0.72	0.38	0.56								
5. Productivity	0.73	0.55	0.65	0.82							
6. Negative Climate	-0.47	-0.14	-0.26	-0.50	-0.54						
7. Instructional Learning Formats	0.56	0.72	0.51	0.34	0.46	-0.17					
8. Content Understanding	0.50	0.59	0.68	0.35	0.47	-0.02	0.71				
9. Analysis and Inquiry	0.23	0.51	0.51	0.10	0.19	0.07	0.54	0.73			
10. Quality of Feedback	0.66	0.51	0.73	0.64	0.61	-0.16	0.53	0.74	0.67		
11. Instructional Dialogue	0.47	0.61	0.78	0.41	0.54	-0.09	0.41	0.67	0.50	0.60	
12. Student Engagement	0.71	0.61	0.54	0.58	0.77	-0.54	0.47	0.39	0.29	0.57	0.39

Note: The numbers 1 to 11 on the columns refer to the dimensions listed in the rows on the left.

Table B.5 displays the correlations between the 12 CLASS dimensions. This serves as a useful check for the internal consistency of the results. As expected, a Positive Climate is positively correlated with Teacher Sensitivity and Regard for Student Perspectives. Productivity is highly correlated with Behavior Management (0.82) and Student Engagement (0.77). Other interesting correlations include Regard for Student Perspectives with Quality of Feedback and Instructional Dialogue.

While not a complete method for validating the use of the CLASS tool, the above correlations do serve as a good guide for potential issues. If, for example, Negative Climate was positively correlated with Student Engagement, there may be reasons to doubt whether classes were being accurately scored. The correlations also point to some potential weak points in the structure of the observed classrooms. Behavior Management, for example, is only weakly correlated with Content Understanding and Analysis and Inquiry. This may indicate that while teachers are managing class environments well, they are not as successful in eliciting a robust interaction with students.

Table B.6: Correlations between CLASS Domains

	Emotional Support	Classroom Organization
Classroom Organization	0.71	
Instructional Support	0.78	0.46

A look at the correlation between CLASS domains reveals an interesting disparity: while Classroom Organization is highly correlated with Emotional Support, it is much less correlated with Instructional Support (table B.6). Even if a teacher is effective at keeping the class organized, that does not necessarily imply that students are being engaged successfully in an instructional manner.

Cronbach's alpha

Cronbach's alpha is a measure of the internal consistency of data. In this case, the alpha is calculated from the component dimensions of each domain, measuring the intercorrelation of the separate dimensions. A higher value indicates higher intercorrelation between the dimensions, and therefore a higher validity of the overall domain.

Table B.7: Cronbach's *alpha*

	Guangdong	S-MTP	MET	UTQ-A
Emotional Support	0.74	0.87	0.87	0.90
Classroom Organization	0.75	0.88	0.90	0.86
Instructional Support	0.85	0.88	0.92	0.91
Overall	0.86			

Note: Cronbach's alpha ranges from 0 to 1. Under 0.5 is generally considered unacceptable, above 0.9 is excellent, and anything above 0.7 can be considered good.

The scores for Guangdong were slightly lower than those achieved by other CLASS studies, although they are still high overall (table B.7).

Summary

The reliability tests of raters and the tool show results in line with comparable assessments, and should be viewed as a positive endorsement of the validity of the tool within the Guangdong context. There are certainly weaknesses that should be remedied with further training and evaluations, but as a starting point, the initial observations are quite promising, especially given the small sample size.

APPENDIX C A PROPOSAL FOR A RESULTS-BASED FINANCING MECHANISM

C.1 Objectives of the pilot in the context of results-based financing

The project design team had discussions with the government on how best to use classroom observation tools. One avenue under discussion was the use of classroom observation techniques to incentivize training providers responsible for developing and delivering in-service professional development courses to teachers. The pilot of the CLASS instrument was conducted with a view to the sequence of activities that a results-based mechanism might entail:

a. assess teacher practices and identify areas in need of improvement
b. design a system that would reward training content developers based on the effectiveness of their programs, based (at least in part) on measuring changes in teaching practices
c. work with training content developers to provide enhanced training that addresses identified areas in need of improvement
d. reassess teacher practices and reward training providers that are able to develop and deliver training courses that lead to meaningful changes in teaching practice.

The ultimate objective was to help determine whether and how classroom observation results could be used to measure the impact of teacher training. This has been done with an eye toward developing a results-based payment modality for training providers that incentivizes the development and delivery of training that meaningfully changes teaching practices and which recognizes results through rewards.[7]

C.2 Results of the pilot

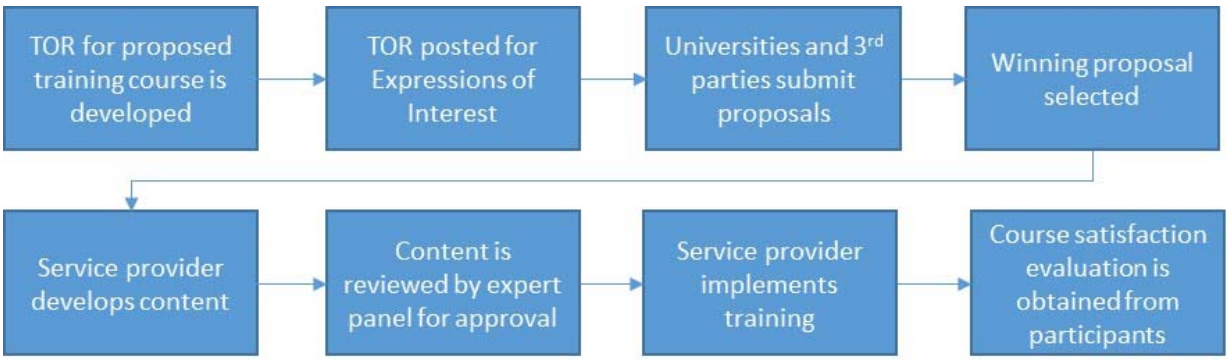
The pilot of the lesson observation instrument garnered acceptance and enthusiasm in use of the CLASS tool for the project. Lesson observation is seen as a valid, reliable, and useful way of measuring impact of the training programs to be implemented in the project. It is also seen as a useful input into the Quality Assurance and Monitoring and Evaluation (QAME) component to be implemented under the project. The team explored how an RBF approach could possibly be integrated into the QAME in the longer term.[8] Below is a general design of how the RBF could potentially function.

C.3 Current system of teacher training

The RBF approach would not be a complete redesign of the current process for delivery of teacher training. Rather, it would involve introducing key elements along the existing process chain. Currently, the following general chain exists in Guangdong:

[7] The pilot of the classroom observation instrument is very useful when considering teacher professional development activities because often the desired outcome would relate to changes in the classroom. In some cases, though, classroom observation may not be appropriate. For example, with a professional development activity that focuses purely on subject knowledge (e.g. a course on geometry) it may be more appropriate to use a knowledge assessment to assure that teachers have obtained a minimum threshold of knowledge from the course (or possibly a pre-post with a targeted/specified improvement in knowledge). While the geometry knowledge gained would ideally result in improvements in teaching practices in the classroom, the use of a classroom observation instrument would not be able to detect changes in knowledge.

[8] Discussions since the pilot confirm that the government is not interested in implementing such an approach in the near term.



C.4 Design that would introduce the RBF approach

Taking the existing approach as the base, the following modified approach would be used in an RBF approach. Bold text below highlights the change and italicized text explains how an RBF approach would alter the motivations and approach of the service provider and other key stakeholders:

1. The training requestor defines a terms of reference **that specifies the aim of the training program as well as concrete outcomes with measurable indicators**.
   This pushes the requestor to be clear up-front on the purpose of the program and what they hope to achieve.
   The proposal specifies that the outcomes will be tracked and that a bonus will be received based on the outcome.
   This sends a signal of the importance of the outcome and that the trainer will be evaluated based on the outcome.
2. The potential training providers submit proposals, **including steps that would be taken to ensure the outcomes are met**.
   This means the providers must consider from the start how they would design the program with the outcome in mind and would help the requestor evaluate the proposal based on which one is likely to produce the best result.
3. The winning provider then designs the content of the program, **always keeping in mind that the program will be evaluated based on the change it produces**.
   Having the incentive in mind, the provider will, in theory, be more thoughtful about ensuring the content and how activities within the program will develop the skills in a way that teachers are not only able to effectively apply the practices in their own classroom but also that the teachers believe it is a worthwhile approach.
4. The training provider will consider how to best deliver the training in an effective manner, **including how to ensure the outcomes are met**.
   The training provider should, in theory, ensure their materials are properly designed, but that the training is also effectively given. This would involve proper preparation, ensuring the trainers are prepared and other steps are taken (e.g. that the trainees have received pre-training materials, etc.). Often the training is cascaded down to the county level. With the awareness of the outcome incentive, the training provider should, in theory, be more careful to ensure the training is effectively provided at all levels.
5. ***(New step introduced)*** **A baseline is collected from a sample of course participants.**
   In this case, the focus is on teaching practices and teachers would be filmed to measure key indicators such as Content Understanding, Analysis and Inquiry, and Regard for Student Perspectives.
6. Once the course is complete, in addition to the existing course evaluation, **an end-line assessment on outcomes would be collected.**

There would need to be a decision made on how soon after the training the end-line should be collected and how to ensure the comparability of the lessons observed.

7. **(New step introduced) A bonus would be awarded to the provider based on the outcome of the training.**
   Considerations would be the amount of funding that would be required to properly incentivize the training provider and whether it should be an all-or-nothing vs. a progressive amount based on the extent of the outcome.

C.5 Key areas of consideration

(i) Political feasibility and associated barriers

Introduction of an RBF scheme would have an impact on many key stakeholders. The political feasibility is a key consideration. Would the service providers accept such a radical change in the way payment is received? Would it end up favoring universities or private providers? Would the risks or uncertainty in the amount of payment be too much for service providers and actually have the unintended consequence of discouraging some qualified service providers from applying? Would there be other unintended consequences that may have a negative impact in either the short-term or long-term for individual courses and for the system of professional development as a whole? These are key considerations in general and would vary depending on the design of the system. The subsequent areas of consideration below would need to be considered within the context of impact on political feasibility.

(ii) Technical feasibility and associated barriers

Introduction of measures for the RBF creates a new layer of complexity to the system. It requires the establishment of new measures and, while some measures could be applicable across many types of training, typically there will need to be a unique design for each training program. This is particularly true for an RBF where the stakes are high, and the measures would need to be particularly accurate and sensitive to a given training course in order to be accepted by the service providers.

There are many technical factors that would need to be taken into consideration. The following are just a few that might help to highlight the complexity and challenges:

- The content developer and training provider will oftentimes be the same entity, but when a cascade approach is used in training there will be multiple layers of delivery. This will tend to complicate the RBF approach.
- The measures will need to be extremely accurate in order for the results to stand up to scrutiny. If the measures are questionable and the service providers don't get the RBF incentive based on the result, then it is highly likely that the service providers would protest the result.
- The technical implementation would require specific skills and expertise, which the existing structure likely does not have, or the qualified personnel would not have sufficient time to dedicate to the measure.

(iii) Balance in base payment vs. RBF incentive

When considering the RBF incentive, a key consideration is what would be required for adequate motivation in the RBF component vs. what would be required in the base payment component to reduce risk to acceptable levels. If the RBF incentive makes up a large proportion of the overall payment, the service providers are at risk of coming out behind financially. This would likely discourage service providers from applying for training programs. If, on the other hand, the RBF incentive makes up a very small proportion of the overall payment then the service providers may not consider the outcomes to be a high priority. The RBF incentive must be high enough to send

the correct signal of the importance of the outcome while not putting the service providers at too much risk financially. If the RBF were to be 10 percent of the total payment it would likely provide enough incentive while not presenting too much of a risk financially if they failed to achieve the targeted outcome. A more rigorous analysis would need to be determined based on specific training programs before coming to a final number.

(iv) Acceptance of teaching practices as a measure

The RBF could be based on different measures as discussed in the next area of consideration below. Any measure would need to be considered valid and reliable by the service providers since their payment would be based on the measure. They would also need to believe that they have sufficient control over affecting the measure through their service delivery, including the content they develop and the implementation of the training.

Teaching practices are an outcome that the service providers have only partial control over. They could conduct training that is of sufficient quality to instill the necessary knowledge and skills for the teacher, but it is ultimately up to the teacher whether she or he actually implements in the classroom what is learned.

There is also an issue of whether the teacher implements what is learned on the day or days of the observation. If a single lesson is observed, then there is less likelihood of detecting the change than if multiple lessons are observed. But there are costs involved in the observation. In the pilot, two full lessons were observed for each teacher and each lesson was made up of two segments.

In addition to the number of lessons observed for individual teachers, the sample size would also need to be sufficiently large to ensure an accurate and statistically significant result. This could be the Achilles heel of using lesson observation in an RBF scheme. For large-scale and prominent professional development activities, it may be worth the effort and cost to implement a lesson observation measure, but for smaller programs it would be very difficult to justify lesson observation to detect changes in the teaching practices. The cost considerations are explored in more detail later.

(v) Cost-benefit considerations of using RBF in professional development

Implementation of an RBF scheme to service providers for professional development activities will entail additional costs. The extent of these costs will vary depending on the type of professional development and the extent and rigor of measurement, as well as the RBF incentive itself.
The cost of implementing an RBF scheme would be dependent on several factors, including:

- **RBF as a top-up incentive vs. reducing the base payment:** Is the RBF bonus incentive a top-up to what would otherwise be offered or is the base payment reduced so that eventual overall payment (made up of base payment + bonus incentive) is equal to what would otherwise have been offered in the non-RBF scheme?
- **Cost of measures:** Because the RBF requires a measure or measures upon which the bonus would be provided, there would need to be additional activities for data collection. It could be argued that outcomes of training programs should be measured either with or without an RBF to ensure quality of professional development programs. This is in part what the Guangdong project's QAME component would do. In this case, the cost would not be additional (or at the most it would be the additional cost required to have the extra level of rigor required upon which the bonus could be based).
- **Different costs depending on the type of measure:** Each type of measure will have its own costs. A lesson observation measure is among the most expensive, particularly if it involves capturing video as opposed to real-time observation. The use of tests to check

knowledge would be less expensive, but quality tests require expertise and would need to be done by an independent party that would need to learn the course objectives and come up with an accurate and reliable tool. In the Guangdong context, this could possibly be the Teacher Research Institute or a completely independent private organization.

- **Independent evaluation group:** In the case of course satisfaction surveys, the service provider simply hands out an evaluation form at the end of the course. There is almost no cost involved. The stakes involved in the results are also low, so there is little need to have an independent evaluation. In the case of an RBF, the evaluation would need to be done by an independent group. This could be done by the group originally responsible for proposing the training (which, in Guangdong, is the department of in-service teacher training) through the formation of expert panels, but would most likely be best carried out through the hiring of a private and fully independent evaluation group.

C.6 Conclusions on feasibility

The feasibility of implementing an RBF scheme for delivery of teacher professional development activities can be considered both at a general level and with the use of classroom observation as a tool to measure outcomes.

RBF to service providers of professional development

Taking the above points into account, there are many potential obstacles to introducing RBF into the service provider scheme for professional development. It would most likely be extremely difficult to use RBF for all forms of training—particularly small training programs—but could be beneficial and cost-effective for the most critical large-scale and training programs.

Videotaped lesson observation as a measure in an RBF scheme

The use of videotaped lesson observation for RBF is likely too costly to be used in a high-stakes endeavor such as RBF. The sample would need to be quite large in order to achieve sufficient reliability for changes in teaching practices. This could make cost of the measure too high to justify. It may be possible to include in-person lesson observation as a lower-cost alternative.

APPENDIX D GUIDELINES AND PROTOCOLS FOR LESSON OBSERVATION USING VIDEO IN GUANGZHOU

D.1 Purpose

The goal of this pilot study is to demonstrate how videotaping of lessons can be used to better understand what takes place in Guangzhou's classrooms and what practices teachers use. This document provides brief guidelines and protocols for the filming of classrooms. It includes general steps to take in the school visit as well as protocols on how to conduct the actual filming. The protocols are adapted from the TIMSS-R Video Study.[9]

D.2 General Steps for the School Visit

Arrival at the school

Arrive at the school at least one hour before the scheduled shooting. Late arrival can create difficulties in the preparations for the filming. The teacher will have a set time for his or her lesson and the set-up should not alter the lesson time.

First meet with school officials. You should never go directly to the teacher's classroom. Always go to the main office first and meet with the principal or the person who has been assigned as your official contact person.

Once in the Classroom

As soon as you get to the classroom where you will shoot the lesson, two factors will help you determine where to position the camera: 1) information about what will happen during the lesson, and 2) the physical arrangement of the classroom.

Ask the Teacher about the Lesson

Try to find out from the teacher about what will happen in the lesson. Often there will be little time for you to talk to the teacher because even though you arrive early, he or she might be busy teaching. However, if you have a chance, ask the teacher and find out as much as possible about the lesson.

The information you want to find out is:

- Roughly how long the lesson will last (longer than 60 minutes?)
- General outline of the activities of the teacher and students that will take place during the lesson
- Whether the chalkboard will be used
- Which chalkboard, if there are more than one, will be used
- Whether AV materials will be used and where they will be placed.

Choose Camera Position

Camera positioning is documented in detail later in the section *Placement of the* Camera. General rules include the following:

- Try to set up the cameras with the windows at your back, thereby avoiding back light problems.
- Close windows, doors, and blinds as needed to adjust the light and reduce noise. *Make sure the teacher does not mind before you do any of this.*
- If your video equipment requires electricity, find the location of the nearest electrical outlet. If at all possible you will want to plug your camera into this outlet so that in the unlikely event of a battery failure you can use electricity as a backup while you replace the battery.

[9] TIMSS-R Video Study Data Collection Manual, Lesson Lab Inc. of Los Angeles, California

- Move student desks as needed to set up the cameras. We are interested in how desks in classrooms are arranged so you should not ask the teacher to significantly alter the lay out of the classroom. However, it is fine to move a few desks over in order to have better visibility. *Before you move any desks make sure that the teacher does not mind.*

D.3 Documenting Lessons

Classrooms are complex environments where many things are occurring at once. In this section, we describe which aspects of the classroom environment we want to document. We discuss how to locate the camera in the classroom and present some general rules, as well as some specific guidelines, designed to help the camera operator make decisions about where to point the camera and how to shoot the action being documented.

Shooting in Real Time

Because we want to see each lesson in its entirety, all videotaping will be done in real time. The camera will be turned on at the beginning of the class and not turned off until the lesson is over. This means that we can study the duration of classroom activities by measuring their length on the videotape. Obviously, this would not be possible if there are any gaps in the recording. *The tapes will not be edited, but viewed from beginning to end in real time. This means that you must attend to what is being captured on the tape at all times. Nothing will be deleted.*

What to Document

Classroom lessons are complex. What kinds of things need to be captured in the videotape? To answer this question, imagine you are an observer. You walk into the classroom to see what is going on. What do you look at? You cannot look at everything; decisions must be made from moment to moment about what to include and what to leave out. When you are in a classroom observing the lesson and trying to understand what is happening, you will probably attend to three things: the teacher, the students, and the tasks. These are the three things we want you to document, but with a primary focus on the teacher.

Document the Teacher

During the lesson, teachers engage in a variety of activities. For example, they explain concepts and procedures, pose problems, assign tasks, ask questions, write information on the chalkboard, walk around the classroom and assist individual students, etc.

Because the main goal of this pilot is to study teaching practices, it is necessary that we thoroughly and carefully document the teacher's activities and behaviors during the lesson. Make sure that you capture what the teacher is doing, what he/she is saying, and what information he/she is presenting to the class.

Document the Students

When you are observing a lesson in a classroom, you would not only look at the teacher all the time but look at the students as well so that you understand what goes on in the classroom. Make sure that you capture what students are doing and saying during the whole-class interaction, when they are working in groups and on their own. Focus mainly on the activities and behaviors of the students who are interacting with the teacher, but turn to other students as well from time to time because students might be doing different things when the teacher is and is not with them. Of course, you cannot document everything that every student says and does. The goal is to sample student behavior so that what is portrayed in the videotape is representative of what actually happened in the lesson.

Document the Tasks

During mathematics and science lessons, teachers assign various tasks to students. Normally the teacher presents the task to students clearly enough that students understand what they are supposed to do, and it is usually not hard to see in the video what the task is. This is not always the case, however. If the task is

ambiguous or poorly described, many students will be uncertain how to proceed. Or, if the class is broken into small groups, each group may be working on a different task.

In all cases, what we want to see on the video is the task that students are actually engaged in doing, whether or not it is what the teacher intended. To see clearly what students are doing it is often necessary to zoom in close enough to capture what at least a few of the students are working on. Make sure you document how students are actually doing the assigned tasks.

Placement of the Camera

For the purpose of this pilot study only one video camera will be used.[10] The videographer will operate the camera. It should be placed on a tripod, but will also be removed from the tripod whenever it is necessary to document specific aspects of the lesson. It will generally be placed between 1/3 and 1/2 of the way back from the front of the class, and the majority of the time will focus on the teacher and his or her zone of interaction.

The physical arrangement of classrooms and the activities that take place within them vary greatly. The videographer must decide where to place the cameras so that the documentation requirements outlined above can be met to the greatest possible extent. It is helpful, if possible, to talk with the teacher before the class begins to find out generally what is going to happen, and where the action will take place. The camera should be placed so that it can easily tape the main chalkboard or audiovisual device, the teacher, and some of the students in a single master shot. The position should also allow for easy panning to other areas of the classroom.

Rationale for Camera Placement

It is not possible, due to varying classroom configurations, to define a single best position for the camera. However, other video studies have found that placing the camera along the side, 1/3 to 1/2 way back, works best in most classrooms.

This position allows good views of the board in medium and close-up shots, as well as good shots of the teacher's and students' faces in a wide master shot. This position also allows for quick panning to the front and rear of the room as well as an ideal view of the opposite side of the room especially if there is a supplementary chalkboard in that location.

Why not set up in the rear of the room? Although setting up in the rear of the room offers a good view of the entire classroom it also has two major disadvantages. The students are only seen from behind, and the camera will most likely have to zoom in to frame the front of the room, which will tend to accentuate camera movement.

Why not set up in the front of the room? Setting up too close to the front of the room results in oblique angles that make it difficult to see what the teacher is doing and to read the board.

Light Sources

If possible, the camera should be set up on the same side of the classroom as the largest set of windows, thus keeping the major light source at the camera operator's back. This orientation will minimize overexposure due to backlighting. This position also allows a good view of the supplementary chalkboard that is often on the opposite wall from the windows.

If the classroom has windows on both sides of the room, choose the side that looks best overall. Be sure to maintain, however, careful manual exposure of the foreground. In any case, the camera's exposure should be set to manual and adjusted according to the situation.

[10] In the future a more complex two-camera strategy may be used where a second camera is placed in front of the classroom, set to the widest shot possible, and used to capture as many students in the classroom as possible.

Also keep in mind that it often is possible to pull window shades if you feel positioning the camera opposite the windows would be a better alternative. In fact, often you will need to pull the blinds even if the windows are behind you so as to avoid reflection on the board or other equipment.

When to Use Tripod versus Hand-Held

Whole Classwork: It is preferable to keep the camera on the tripod during periods of whole classwork (when the teacher and/or a student is at the board). Circulating through the classroom can be distracting and can make the camera the center of attention.

Independent Work: If independent seatwork occurs for more than 2-3 minutes, it is preferable to handhold the camera, so that you can more closely capture individual interactions and students' work. Below are some examples of seatwork:

- When the teacher assists students individually or as a small group.
- When students break into groups and work on assigned tasks.
- When students gather to work around a computer

During these activities, take the camera off the tripod, handhold it, and walk around the room for the duration of the seatwork period. Try to remove the camera smoothly from the tripod so that you will not lose the action you are documenting. Then, at the end of the seatwork period, put the camera back on the tripod as smoothly as possible.

Keeping track of the teacher. It is very important to keep track of the teacher during periods of independent work. If the teacher is interacting with students, it is particularly important to capture these interactions. Most likely, during periods of independent seatwork you will be filming the teacher interacting with students, using a medium or wide shot. However, you should also try to periodically capture students' work, including what they have written on their paper, the materials they are using, and their textbooks.

Finding opportunities to shoot students' work. Shooting students' work means getting close-ups that are readable to a viewer. Try to get at least one good shot of a student's work. Ideally try to shoot as many different students' work as possible, without losing track of the teacher. One opportunity for shooting students' work is when you are filming the teacher providing assistance. Another opportunity is when the teacher is not doing anything, and you see a student whose work you could easily shoot.

The close-up shot. Getting a good close-up shot of students' work presents a somewhat difficult situation for the videographer. In order for the shot to be effective, the viewer must be able to read what the student has written. Such shots are critical for the viewer to know exactly what work students are doing, or have done, at their seats. However, getting this kind of shot can be disruptive to the student. Therefore, you will have to use your judgment as to when it is appropriate to attempt these close-ups.

For the ideal close-up shot, you should stand behind the student (or possibly to their side), zoom in, focus carefully, and film everything they have written. Please be aware that the camera only needs to be in this position long enough to zoom in and focus, because viewers can easily freeze this frame of video.

Other Issues to Consider in Placing the Cameras

There are still some other issues you will need to consider when choosing the camera positions in a classroom.

- Overhead projectors, slides, multiple AV presentations: You should take into account the audiovisual materials that will be used so as to position yourself at a vantage point from which you can best capture see them.
- Direction in which students are facing: Try to position the camera so that you can see the faces of at least some of the children (if not the majority). This will reduce the chance that you have to remove the camera from the tripod.

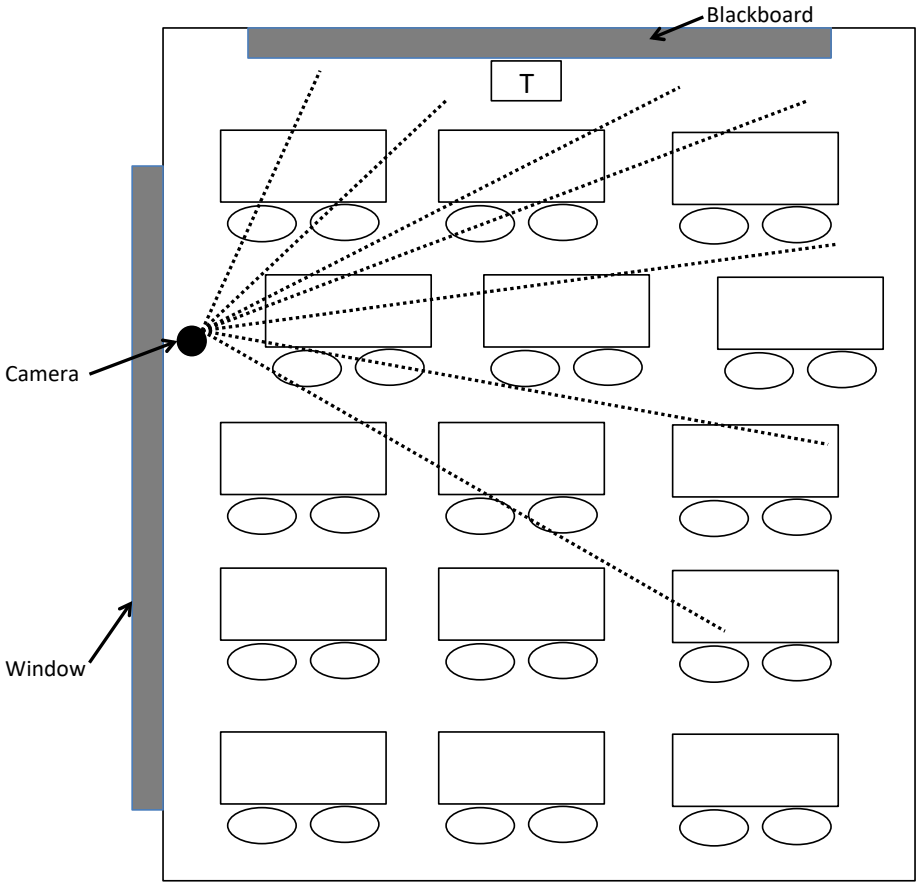
- Clear view: You want to avoid having students sitting directly in front of the camera because they will block your view. If you find a very good position but a student is in your way, you might want to consider asking the teacher if it is OK to ask the student to move.

Some Common Situations, and Where to Place the Cameras

The following section illustrates where to place cameras in a variety of classroom settings with different instructional activities. In general you may find mathematics lessons easier to videotape than science lessons because science lessons are often held in a lab, which tends to be much larger than a regular classroom, and desks are often built-in so that you cannot move them to secure the camera positions. Also science lessons involve demonstrations and experiments that often require a videographer to handhold the camera and move around in the room to document what the teacher and students are doing. In any event, you should always keep in mind in making your decisions of where to place cameras and what to videotape the principles and guidelines described above.

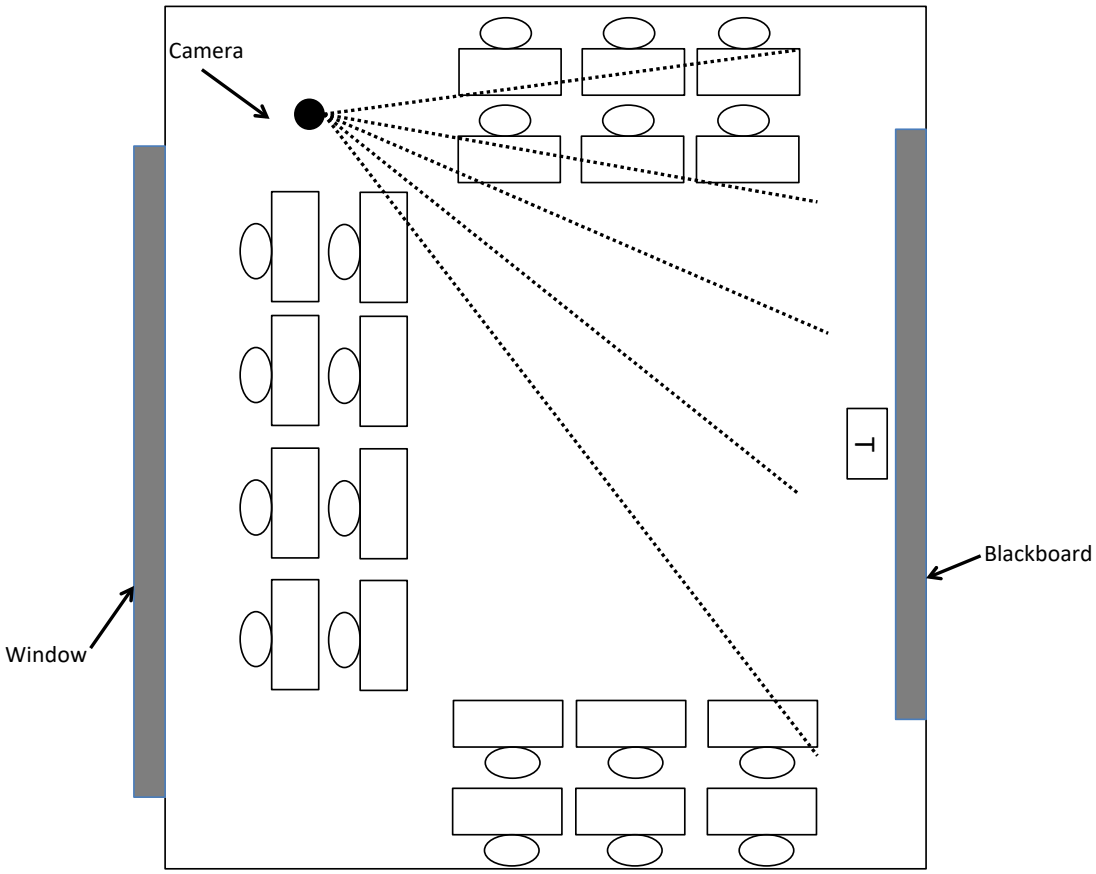
Situation 1: Window Opposite from the Door, Chalkboard at the Front, Movable Student-Desks Facing the Front

This situation is probably the most common classroom setting. You can place the camera by the window, 1/3 of the way from the front, leaving it aimed at the students behind the camera. Keep the camera on the tripod as long as you can document what the teacher and students are doing.



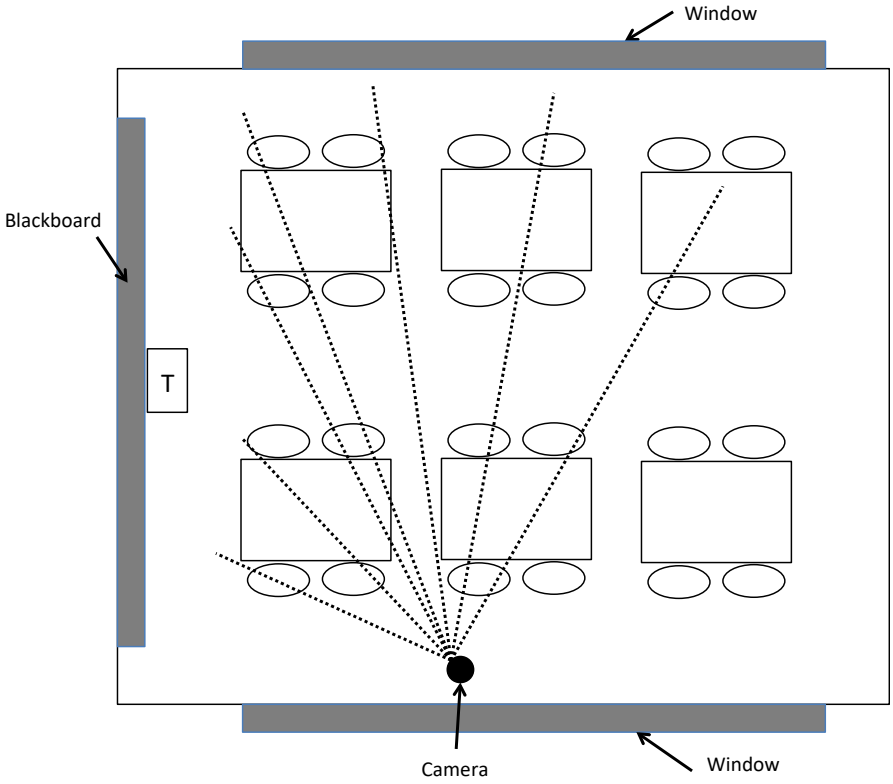
Situation 2: Chalkboard at the Front, Window on the One Side, Student Desks Arranged in a U-Shape

This situation does not allow you to apply the 1/3 view rule. You should place the camera where you have a good view of the teacher and the chalkboard, and students are not blocking your view.



Situation 3: Students Sit in Groups, Windows on Two Sides of the Room

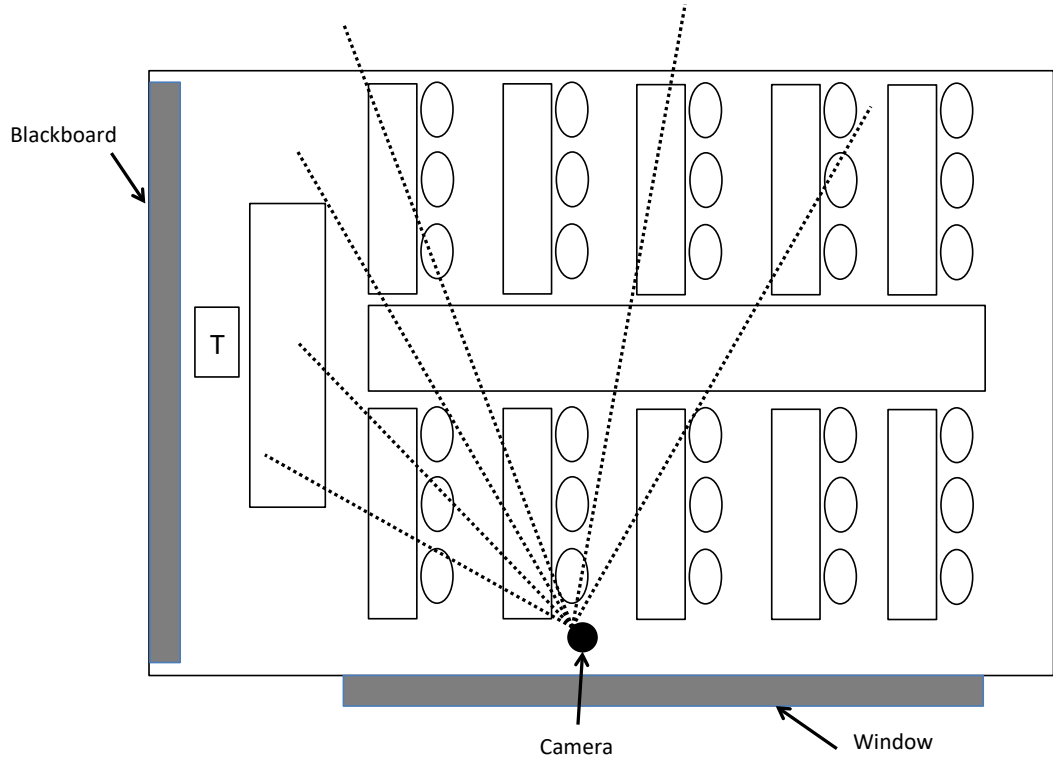
In this situation 1/3 view may apply. Again place the camera so that you have a good view of the teacher. Try to avoid backlighting situation.



Situation 4: Large Science Lab, Student-Desks Not Movable

Often science labs are much larger than normal classrooms, and student-desks are built-in so that you cannot move them. Because the room is large, often there are enough rooms for students to sit even if you occupy

few seat spaces (see the diagram below). However, make sure you ask the teacher if it does not cause any problem.



Deciding Where to Point the Camera

Once you begin videotaping classroom lessons you will see that it is not enough simply to know what needs to be documented. It is impossible to simultaneously capture teacher, students, and tasks. You must decide, at any given moment, where to point the camera, and what to include in the shot.

Keep Track of the Teacher at All Times

Because the teacher is an extremely important part of the lesson, we want to keep track of the teacher at all times. This does not mean, however, that you must always have the teacher in the camera view. We will have audio coming from the teacher's wireless microphone, so as long as you pan back to the teacher frequently we will be able to find out what the teacher is doing. If the teacher engages in a long interaction with a single student we want to capture it, but also we want to see what the other students in the class are doing.

Some Difficult Situations and Their Solutions

In general, you find it difficult to decide where to point the camera when: 1) separate activities are occurring at the same time, and 2) when events change very quickly. Here are the rules to keep in mind when you encounter those situations:

When Separate Activities are Occurring at the Same Time

If students are working on their own on an assigned task while the teacher prepares materials on the board, it is difficult to document both what the teacher is doing and what the students are working on.

The general solution to this situation is to focus on the teacher for a while, then pan slowly away from the teacher to document what the students are doing. It may be necessary to zoom in to see the task that students are working on. Then move back to the teacher.

Rule: Keep the shot mainly on the teacher but tape students activity from time to time to understand what task they are working on.

When Events Change Very Quickly

Sometimes things change constantly during the lesson. You must listen carefully to what is happening and try to predict what might happen next. This is the only way to be ready to react in time. However sometimes changes will occur very quickly and you are likely to miss what has happened.

If events change quickly and it is clear that the change is only a brief one, it is often impossible to catch the change in time and it is better to let it go. In general you should avoid moving the camera to capture brief events. We are likely to miss them anyway, and rapid moves compromise the quality of our tapes. It is not only the brief event that is missed, but parts of the more enduring event would be missed as well as you try to find your way back to the original event.

Rule: avoid moving the camera to capture brief events.

The table below highlights some difficult situations likely to occur in classrooms, what to do when they occur, and why.

	Descriptions of possible situations	**What to do**	**Why**
1	• Teacher at the front talking • One student is at the board working on a problem and talking publicly • Rest of the class working individually at their seats	Focus on the teacher and the student at the board, but find a chance to document what other students are doing	Because we want to document: 1) the teacher, 2) teacher- student interaction, 3) new information on the board, and 4) students' task
2	• Teacher walks around assisting the students privately and talks to the whole class from time to time • One student at the board working on a problem • Rest of the class working individually	Document how the teacher instructs individual students, but document the student at the board and the information on the board when there is a chance	Because we want to document: 1) the teacher, 2) new information on the board, and 3) students' task

3	• Teacher stays at the teacher desk assisting students privately • Rest of the class working on their own	Document how the teacher instructs individual students (move close to them) and document what other students are doing	Because we want to document: 1) the teacher, 2) teacher- student interaction, and 3) students' task
4	• Every group works on the same task; • Teacher walks around and assists each group	Document how the teacher assists individual groups (follow the teacher) and also document some groups when teacher is not with them	Because we want to document: 1) the teacher, 2) teacher- student interaction, and 3) students' task
5	• Every group works on different tasks; • Teacher walks around and assists each group	Document how the teacher assists each individual group (follow the teacher) and also document every different group work	Because we want to document: 1) the teacher, 2) teacher- student interaction, and 3) students' task
6	• Every group works on a different task, • One group works outside the classroom • Teacher walks around and assists each group	Same as #5 but find a chance to document the group outside	Because we want to document: 1) the teacher, 2) teacher- student interaction, and 3) students' task
7	Whole class leaves the classroom and work outside	Follow the class and videotape outside	Because we want to document: 1) the teacher, 2) teacher- student interaction, and 3) students' task

Audio Recording

Use a dedicated microphone on the teacher if possible. Capturing the voice of the teacher is critical. While a dedicated microphone on a video camera will often be sufficient when a teacher is lecturing, it becomes much more difficult to capture the teacher's voice during group work when many students are speaking. Having a dedicated microphone on the teacher is very beneficial for capturing what the teacher is saying at all times. It is also helpful for conversations between the teacher and students during group work sessions where the teacher is going from one group to another.