Public Disclosure Authorized

Public Disclosure Authorized

CHINA

Can Classroom Observations Measure Improvements in Teaching?

FEBRUARY 2018

REACH funded a pilot of the Classroom Assessment Scoring System (CLASS) to test its usefulness as a tool to assess teaching practices that can help inform the design of incentives for teacher training providers.

The Results in Education for All Children (REACH) Trust Fund *supports and disseminates research on the impact of results-based financing on learning outcomes. The EVIDENCE series highlights REACH grants around the world to provide empirical evidence and operational lessons helpful in the design and implementation of successful performance-based programs.*

Student test scores alone are not an accurate measure of teacher quality.

Systematic observations of teaching practices can be a useful tool to understanding what makes a good teacher.

While countries around the world have made enormous strides in terms of increasing access to education, many education systems are now working to ensure the high quality of learning for all students. There are two measurement challenges involved in the pursuit of this goal—how to measure student learning outcomes and how to measure teachers' role in achieving those outcomes. While student test scores can be used to measure teacher quality, they miss a lot of what comprises effective teaching. Systematic observations of teaching practices can be a useful tool for shining a light on the "black box" of what makes a good teacher and can help policymakers to design better curricula and teacher training programs.

Furthermore, a consistent system for assessing teaching quality is a precondition for using results-based financing (RBF) schemes that make financing conditional on improvements in teaching quality. RBF can be used as an incentive to foster improvements at the level of the teacher, school, school system,

This note was adapted from Coflan, Andrew, Andrew Ragatz, Amer Hasan, and Yilin Pan (2018). Understanding effective teaching practices in Chinese classrooms: evidence from primary and junior secondary schools in Guangdong, Policy Research Working Paper, World Bank, Washington D.C.

or teacher training provider. However, the process of linking financing to learning outcomes is often plagued by concerns about cheating, "teaching to the test," and exams that do not fully reflect what students have learned. Ensuring that RBF works effectively in the Chinese context will depend on developing ways to measure teaching quality that can be influenced by teachers, teacher trainers, and other actors but not manipulated.

The Results in Education for All Children (REACH) Trust Fund at the World Bank has funded a pilot of the Classroom Assessment Scoring System (CLASS) in Guangdong, China to test its usefulness as a tool to assess teaching practices. The pilot was also designed to establish a proof of concept for using classroom observations to measure the impact of teacher training and incentivize training providers within an RBF mechanism. In the pilot, the CLASS tool was used to conduct classroom observations of 36 teachers in Guangdong and to assess the strengths and weaknesses of their teaching practices. It sought to test whether this tool could be used to measure teaching practices in the context of China and how to introduce classroom observations into a quality assurance and monitoring and evaluation (QAME) system, as well as exploring how to establish the preconditions for introducing RBF into China.

The CLASS tool provided valid and reliable measures of teaching quality in Guangdong as it has done in other settings. On average, teachers in Guangdong scored high on classroom organization but lower on emotional support and instructional support. This was useful evidence showing how the teacher training system can fill these critical gaps in teaching practices. While there was substantial variation in performance among teachers, there was only modest variation by county, between urban and rural areas, by teacher type, by grade, or by years of experience. However, teachers with more student-centered beliefs scored significantly higher than those with teacher-centered beliefs.

This pilot has also yielded lessons about how to improve the design of classroom observations in the future to be more effective in an RBF context. To address discrepancies between the ratings of different classroom observers, it may be necessary to provide them additional training, or to modify the scoring rubric to take cultural issues into consideration. In designing an RBF mechanism using classroom observations, it is critical to design incentives targeted to teaching practices where there is significant room for improvement.

Guangdong

Schools in rural areas are often under-resourced and staffed by teachers of variable quality.

CONTEXT

The issue of ensuring that all students receive high-quality and equitable learning is particularly acute in China. While Guangdong's GDP has grown by 13 percent per year since 1981, this rapid economic progress has coincided with growing urban-rural disparities. The urban-rural income gap increased from 1.7 in 1980 to over 3.0 in 2010.[1] Historically, the system for financing public services has resulted in substantial spending disparities across regions, particularly in the education sector, resulting in insufficient infrastructure, faculty, and operational resources.[2] While enrollment rates are close to 100 percent thanks to large investments in public schooling, schools in rural areas are often under-resourced and are staffed by teachers of variable quality.[3] These disparities also translate into gaps in learning performance. In 2015, results from four provinces (Beijing, Shanghai, Jiangsu, and Guangdong) on the Program for International Student Assessment (PISA) found that rural students lagged roughly 2.5 years of learning behind urban students.[4]

To reduce social inequality and achieve a more harmonious society, one of China's key priorities is to ensure equity in both the quantity and

quality of education.[5] To reach this ambitious goal, China must continue to invest in teacher quality, including increasing teacher training capacity. Training in Guangdong is typically provided by teacher training colleges but varies in quality from both a content and delivery standpoint. There is no consistent system for measuring training objectives. As a result, with no way to measure whether their training is meeting expected outcomes, teacher trainers are often unaware of how to improve their content or delivery method.

Achieving equity will also require a quality assurance mechanism, which is currently lacking. While the province measures some overall indicators of the education system, such as the number of backbone teachers (individuals that have been identified as excellent teachers by their peers; they also receive additional professional development), the overall system does not include any metrics of teaching quality in the classroom. Guangdong is in the process of improving its QAME system as part of the Guangdong Compulsory Education Project, a World Bank co-financed initiative to improve education in 16 under-performing counties in the province. These 16 project counties have significantly lower per capita GDP, higher levels of poverty, larger rural populations, and higher reliance on agriculture than the province overall. These 16 counties also lag behind in meeting Guangdong's "Chuang Qiang" standards for school quality. This study focuses on three of the 16 project counties, Wuhua, Dianbai, and Lianjiang, one low-performing, one medium-performing, and one high-performing county in terms of education performance.

WHY WAS THE INTERVENTION CHOSEN?

In a recent literature review, pedagogical interventions and teacher training were found to be among the most effective methods to improve student learning outcomes.[6] However, changing teaching practices is notoriously difficult, and these efforts tend to fail if not well designed or properly targeted. A recent study found that, while teachers in China gained knowledge from training, their teaching behavior did not change, leading to no significant gains in learning.[7]

Furthermore, measuring classroom teaching practices consistently is a key challenge in ensuring teacher quality. While in Guangdong, annual grade-wide examinations are used effectively to measure learning outcomes, attempts to measure the effect of teaching practices on learning outcomes have been less effective. Teachers in Guangdong are often observed in the classroom, but these observations are subjective and idiosyncratic and do not measure quality in a standardized way. Without a consistent metric of instructional quality, it is difficult to assess whether teaching is improving over time or whether teacher training is effective. Teacher training courses are not designed to address observed weaknesses of teacher practices, and there is little focus on measuring training outcomes. In the absence of any standardized measurement of teaching quality, there is little incentive for teachers to improve or for teacher training providers to develop content that will improve teaching practices. Before implementing RBF, school administrators must be able to demonstrate to teachers or trainers that there is a need for them to improve and in what areas.

Classroom observations accompanied by detailed feedback on each teacher has emerged as one of the most promising ways of improving teaching and learning. The CLASS tool was chosen for this study because it has been implemented in a number of countries and proven to be one of the most valid and reliable

classroom observation instruments. It has also been shown to be particularly effective in producing teacher assessment results that are correlated with student learning outcomes, as measured by standardized test scores in the U.S.[8] and in developing country settings.[9]

A pilot of the CLASS tool was conducted in Guangdong as a proof of concept with three objectives: (a) to assess to what extent an internationally validated measure of teaching practices could be applied in the Chinese context; (b) to gain an understanding of how to use classroom observations to assess the strengths and weaknesses of existing teaching practices in order to improve teacher training and monitoring and evaluation (M&E); and (c) to establish the preconditions for an RBF mechanism for teacher trainers. The ultimate goal was to establish a reliable measure of teacher quality agreed on by all relevant stakeholders so that it could be used to assess how a teacher performed before and after training and to reward trainers with performance-based bonuses.

HOW DID THE INTERVENTION WORK?

CLASS is a classroom observation tool that measures the quality of teacher-student interactions, which are the main mechanism through which children learn. It was developed by researchers at the University of Virginia to provide an objective, quantitative measurement of teaching practices. Today, CLASS observation training is administered by Teachstone, a private U.S.-based company. As of April 2016, the tool had been validated and implemented in a number of countries. While such validation is always necessary when bringing a tool into a new setting, it is especially critical for China given that the CLASS tool is designed for a student-centered learning environment, while Chinese classrooms are historically more traditional and teacher-centered, focused on lecturing and the transmission of knowledge. Furthermore, before the government adopted the CLASS tool to determine incentives for teacher trainers, it was critical to demonstrate that it was reliable and accurate when applied in Chinese classrooms.

Table 1. CLASS Dimensions

CLASS domain	CLASS dimension
Emotional support	Positive climate
	Negative climate
	Teacher sensitivity
	Regard for student perspective
Classroom organization	Behavior management
	Productivity
	Instructional learning formats
Instructional support	Content understanding
	Quality of feedback
	Analysis and inquiry
	Instructional dialogue
Student engagement	Student engagement

In other countries, the CLASS tool has been used in all grades from kindergarten to secondary school. In this pilot, it was tested in primary schools (grade four) and junior secondary schools (grade eight). While the tool is subject-agnostic, in this study it was used to observe English, math, and Chinese classes. In each classroom, two raters used the tool to score classroom behavior on a 1 to 7 scale in 12 dimensions, organized into four domains (*Table 1*).

In addition to the classroom observations, teachers were given a questionnaire to fill in to assess their beliefs and attitudes about teaching. This questionnaire consisted of questions on topics such as the role of teachers in the classroom, the role of assessments, and the structure of student-teacher interactions. Its purpose was to measure each teacher's alignment with either student-centered or teacher-centered beliefs.

The study sample consisted of 36 teachers in 12 schools, with each of the three pilot counties each contributing two primary schools (urban and rural) and two junior secondary schools (urban and rural). Each school included one new teacher (with less than three years of experience), one backbone teacher, and one potential backbone teacher (who has not yet completed backbone teacher training). Within each type of school, the teachers were also evenly distributed between English, math, and Chinese. This breakdown was designed to assess the CLASS tool's validity across a balanced

distribution of teachers and to assess teacher performance across several relevant dimensions, including teacher experience, school level, urban vs. rural, and school subject.

CLASS includes four cycles of 15-minute observations of teachers and students by certified observers. CLASS has been validated in over 2,000 schools, primarily in the United States.

WHAT WERE THE RESULTS?

The results of the pilot showed that teachers in Guangdong have strong classroom organization skills but score lower on emotional support and instructional support, which is similar to results that have been seen in the U.S. Teachers scored very high on classroom organization, with average scores of 6.5 (out of 7) for primary school teachers and 6.2 for junior secondary school teachers (*Table 2*). Scores of 6 out of 7 mean that most teachers are able to prevent and redirect students' misbehavior, manage instructional time, and maximize students' engagement effectively. In contrast, primary and junior secondary teachers averaged 4.2 and 3.7 respectively on emotional support, and 3.7 and 3.6 respectively on instructional support. Scores of 4 out of 7 in these areas imply that many teachers are unable to establish emotional connection with students, create a positive learning environment, or promote a depth of understanding in their students. These scores suggest that teachers in China have similar levels of teaching quality and similar strengths to those of U.S. teachers as shown in three studies in the United States (although it is difficult to assess whether such comparisons may be influenced by the raters' pre-existing cultural attitudes).

However, these averages mask a substantial variation in the classroom performance of individual teachers as well as modest variations by county, urban/rural location, teacher type, grade, and years of experience. The breakdown by individual teachers reveals large disparities in performance. For example, in grade eight, their emotional support scores ranged from 1.7 to 5.1, with the best teacher scoring three times higher than the one with the lowest score. However, average classroom performance was similar in each pilot county despite large differences between them in terms of economic prosperity and educational achievement. Similarly, there was only a slight difference in average scores between urban and rural schools, with rural classrooms performing slightly better in all three domains. That being said, even though the "urban" schools were located in the county seats, all schools in the pilot counties could be considered to be rural compared with Guangzhou.

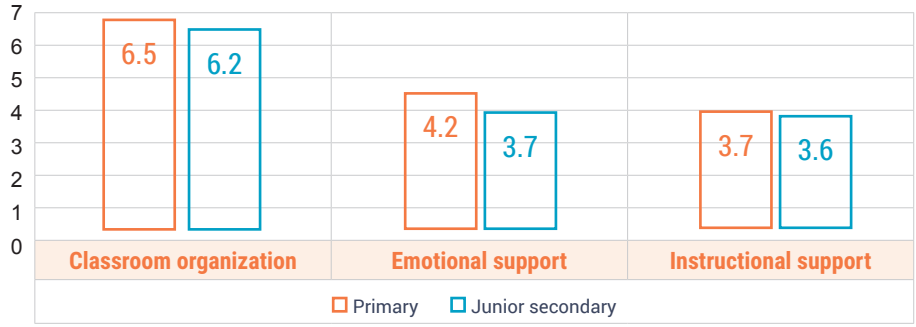
The differences in scores between new teachers, backbone teachers, and potential backbone teachers were also relatively small. Potential backbone teachers scored highest in all three domains, perhaps because they are more likely to be tenured than new teachers, have been identified as strong performers, and are more motivated than backbone teachers to perform well. These gaps were larger in junior secondary schools, suggesting a more difficult transition for new teachers, perhaps due to the larger class sizes, more difficult curricula, and the students' greater emotional needs. Grade four teachers consistently scored higher than their grade eight counterparts in all three domains. The youngest teachers performed best on emotional support and the oldest scored highest on instructional support and classroom organization.

Teachers with more student-centered beliefs scored significantly higher than those with teacher-centered beliefs. On the beliefs survey, all teachers were scored based on their alignment with either student-centered or teacher-centered beliefs. Scores associated with the most teacher-centered beliefs were negatively correlated with CLASS scores in the emotional support, classroom organization, and student engagement domains, while scores associated with student-centered beliefs were positively correlated with all domains and more strongly correlated with instructional support scores. This suggests that a student-centered teaching approach, in addition to fostering a positive classroom environment, is also more conducive to effective instruction and classroom organization.

The CLASS tool provided valid and reliable measures of teaching quality, with high levels of agreement between raters. The percent agreement, defined as the fraction of scores that were equal or adjacent (±1) between the two raters, was 75 percent on average, similar to the 77 to 80 percent agreement in the three comparable U.S. studies. In addition, many of the 12 dimensions were strongly correlated with each other, with an average correlation of 0.51, suggesting that the tool is internally consistent and that teachers tend to perform consistently well (or poorly) across all dimensions. The emotional support scores are strongly correlated with both classroom organization (0.71) and instructional support (0.78). However, there was a lower correlation between classroom organization and instructional support (0.46), which suggests that even teachers who can keep their class organized may not be effectively engaging their students in instruction.

However, there were larger discrepancies between raters on some dimensions for both technical and cultural reasons. Another way of examining differences between multiple raters is to consider the magnitude of the spread between their scores. For example, a spread of 3 or more on a 1 to 7 scale indicates a very serious discrepancy in judgment by one of the two raters. Certain dimensions proved harder for Chinese raters to agree on, including analysis and inquiry, instructional dialogue, regard for student perspectives, and teacher sensitivity, on which the raters scored at least 10 percent of teachers with a spread of 3 or more. In some cases, the challenges were mainly technical and could be addressed by improving the training provided to raters. However, in the cases of regard for student perspectives and teacher sensitivity, cultural factors may explain the discrepancies. These dimensions are not usually considered in existing teacher observations in Guangdong, and therefore raters may have very different opinions about what constitutes these characteristics in China.

Table 2: Average Teacher Scores on Each CLASS Domain



WHAT WERE THE LESSONS LEARNED?

To use RBF to improve results in education, many preconditions must be met. One precondition is establishing a set of indicators that is accurate and reliable, easy to measure, and has a strong causal link to ultimate learning outcomes. The main objective of this pilot was to determine whether the CLASS tool could provide effective indicators to measure improvements in teaching practices over time. If so, this would pave the way to using RBF to incentivize training providers by rewarding them depending on the impact of their training.

While the observations in the pilot were not taken at different points in time, the variation in scores can help to answer this question. The scores in emotional support and instructional support varied widely, with teachers' scores ranging from 1.7 to 5.1 and 1.7 to 4.6 respectively. However, teachers scored very high on classroom organization and student engagement with little variation. Ninety-eight percent of all teachers received a 5 or higher in classroom organization, and 90 percent received a 5 or higher in student engagement. With so little variation, there would be little possibility of capturing measurable changes over time. To measure meaningful improvements in these domains, it may be necessary to modify the scoring rubrics to identify areas where teachers are weak.

The pilot also identified some challenges in the rating of specific dimensions. In some cases, the issues are likely to have been technical and could be addressed by providing raters with more extensive training. In other cases, cultural factors may have caused discrepancies in scores across raters. In these cases, the CLASS scoring rubrics may need to be modified to make them more applicable to China. In fact, it is critical to pay careful attention to adapting any classroom observation tool to the Chinese context to ensure that all relevant stakeholders consider the tool to be valid and culturally relevant and trust its results.

This pilot also examined the implications of observing classrooms either in person or via a video link. There are advantages to each of these options. For example, a video can be scored by many different raters and can be re-watched at any time to resolve any discrepancies, while in person observation is a one-time experience, but generally allows for a better evaluation of the classroom environment and teacher-student interactions. While no definitive conclusions could be drawn from the pilot's analysis of this question, the in-person observations yielded higher scores in most dimensions, but further investigation will be needed to select the best option going forward.

While this pilot was conducted in only three counties in Guangdong and therefore cannot be considered to be representative of all of China or of the province, it shows that the CLASS tool is an accurate, reliable indicator of teacher quality in the Chinese context. The results from the pilot can be used to inform the design of an RBF scheme to establish performance-based contracts for teacher training providers. However, the design of such a program must also meet several additional preconditions, including how to ensure that training providers agree to the financial incentives, how to manage the additional complexity and cost of RBF contracts, and how to provide feedback to teacher trainers to ensure that they have the information that they need to improve how they do their jobs.

The CLASS tool proved to be successful in providing valid and reliable measures of teaching quality in the three pilot counties in Guangdong.

CONCLUSION

In June 2017, at a workshop attended by Department of Education officials, classroom observers, and teachers, there was strong agreement on the tool's usefulness and applicability. Teachers scored high on classroom organization but lower on emotional support and instructional support, which was similar to CLASS results in other settings. This pilot established a baseline of teaching practices in Guangdong and identified strengths and weaknesses that could be taken into account in the design of teacher training and the new QAME system. The pilot has shown that classroom observations can be used as an outcome measure in an RBF scheme to give teacher training providers incentives to change teacher behavior. Ensuring that such a scheme is carefully designed will be critical to ensuring support from all stakeholders and to establishing clear links between training and expected outcomes.

1 Coflan, Andrew, Andrew Ragatz, Amer Hasan, and Yilin Pan (2018). Understanding effective teaching practices in Chinese classrooms: evidence from primary and junior secondary schools in Guangdong, *Policy Research Working Paper*, World Bank, Washington D.C.

2 Tsang, M.C., and Y. Ding. (2005). "Resource utilization and disparities in compulsory education in China." *China Review: An Interdisciplinary Journal on Greater China,* 5(1): 1–31.

3 Wen J. Peng, Elizabeth McNess, Sally Thomas, Xiang Rong Wu, Chong Zhang, Jian Zhong Li, and 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

4 OECD (2016), PISA 2015 Results (Volume I): Excellence and Equity in Education, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264266490-en.

5 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.

6 Evans, D., and A. Popova. (2016). "What really works to improve learning in developing countries? An analysis of divergent findings in systematic reviews." *The World Bank Research Observer*, Volume 31, Issue 2, pages 242–270.

7 Lu, M., P. Loyalka, Y. Shi, F. Chang, C. Liu, and S. Rozelle (2017). The Impact of Teacher Professional Development Programs on Student Achievement in Rural China, *Rural Education Action Program Working Paper 313,* Stanford University, Stanford, CA.

8 Allen, Joseph, Anne Gregory, Amori Mikami, Janetta Lun, Bridget Hamre, and Robert Pianta (2013). "Observations of Effective Teacher–Student Interactions in Secondary School Classrooms: Predicting Student Achievement With the Classroom Assessment Scoring System—Secondary." *School Psychology Review,* 42, 76–98.

9 Araujo, M. Caridad, Pedro Carneiro, Yyannu Cruz-Aguayo, and Norbert Schady (2016). "Teacher Quality and Learning Outcomes in Kindergarten." *The Quarterly Journal of Economics,* Volume 131, Issue 3, 1 August 2016, Pages 1415–1453.

PHOTO CREDITS:

Cover: Project photo courtesy of the World Bank.
Page 3: "China_2010_Peng-Yang" by SIM USA, license: CC BY-SA 2.0
Page 5: Project photo courtesy of the World Bank.
Page 7: "Classroom" by WabbitWanderer, license: CC BY-SA 2.0

RESULTS IN EDUCATION FOR ALL CHILDREN (REACH)

REACH is funded by the Government of Norway through NORAD, the Government of the United States of America through USAID, and the Government of Germany through the Federal Ministry for Economic Cooperation and Development.

worldbank.org/reach
reach@worldbank.org

Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung
Norad
USAID FROM THE AMERICAN PEOPLE
WORLD BANK GROUP Education