ED/GEMR/MRT/2016/P1/30

United Nations Educational, Scientific and Cultural Organization

Global Education Monitoring Report

Background paper prepared for the 2016 Global Education Monitoring Report

Education for people and planet: Creating sustainable futures for all

Measures of quality through classroom observation for the Sustainable Development Goals: Lessons from low-and-middle-income countries

By RTI International

2016

Abstract

With the adoption of the United Nations General Assembly's Sustainable Development Goals (SDGs), global education agencies are grappling with how quality can and should be measured for global reporting purposes. Several factors at the education system, school, and classroom levels shape education quality, including the limited information available at the global level about what is happening in the classroom. Such information can only come through observation-based measures that record teacher practices, either through routine monitoring conducted by system actors or through surveys. Classroom observation is used extensively in not only teacher education and professional development, but also in evaluation studies. However, there are fewer cases where classroom observations are used for system monitoring purposes—particularly in low- and middle-income countries. This paper reviews what has been learned from observation instruments in low- and middle-income countries and what opportunities (i.e., scope) there are to systematize these countries to that they can monitor quality at both the school and system levels.

1 ACKNOWLEDGEMENTS

This paper relies on data and experiences gathered from several US Agency for International Development (USAID) funded programs. Some sections of this paper are drawn from prior reports and are cited accordingly. The authors[1] gratefully acknowledge the support and efforts of USAID, host country governments, staff, teachers and students, without whom this research would not be possible.

[1] Developed by: Sarah Pouezevara, Alison Pflepsen, Lee Nordstrum, Simon King, Amber Gove

Contents

2 List of Figures

3 List of Tables

1. Introduction

Education researchers have long sought to define education quality, from Horace Mann's efforts toward standardization (Kantor and Lowe 2004), to recent initiatives incorporating observational feedback and value-added measures (Harris and Herrington 2015). With the recent adoption by the United Nations General Assembly of the Sustainable Development Goals (SDGs), global education agencies are now turning to the understanding of how to monitor and track all 17 goals and 169 targets across more than 190 countries under the SDG. For education, Targets 4.1 and 4.2 call on countries to 'ensure that all girls and boys complete free, equitable and **quality** primary and secondary education leading to relevant and effective learning outcomes' (United Nations Educational, Scientific and Cultural Organization [UNESCO] Technical Advisory Group 2015) and '…have access to **quality** early childhood development, care and pre-primary education so that they are ready for primary education' (UNESCO Technical Advisory Group 2015).' Of particular interest to this paper is the question of how quality can be measured and monitored for global reporting purposes.

Several factors at the education system, school, and classroom levels shape education quality. Measuring access and achievement at the student level provides information about outcomes, but relatively little information about what processes led to those outcomes and, as such, limited information with which to improve instruction. As a result, there is relatively little knowledge available to global reporting agencies about what happens in a classroom. One way to improve this gap, as noted in a recent education rigorous literature review (Westbrook et al. 2013), is obtaining information through systematic observation-based measures that record teacher practices, either through routine monitoring conducted by system actors (i.e. instructional supervisors or coaches) or through external surveys.

Observational methods are used extensively in teacher education and professional development to describe and evaluate classrooms. Pianta and Hamre (2009) argue that although observation can be a central feature of accountability frameworks, the most important reason to conduct classroom observation is to inform teacher professional development and, subsequently, to know if it is working. Observational methods can also be inquiry-driven, investigating classroom processes in order to generate hypotheses about their impact on learning (Pianta and Hamre 2009). There are fewer cases where classroom observations are used for system monitoring purposes (i.e. designed to gather information that can inform policy and practice of education systems *at scale*), particularly in low- and middle-income countries (LMICs).

This paper reviews what has been learned from classroom and lesson observation instruments in LMICs and considers what opportunities (i.e., scope) there are to systematize these countries to help them monitor quality at both the school and system levels.

2. Effective teaching practices associated with improved learning outcomes

In the education field, it seems intuitive that a classroom teacher exerts the most influence on student learning in any school system. Common thought is that schooling systems are only as good as the teachers that comprise them. However, beneath this nearly ubiquitous valuation of teachers and teaching, research has repeatedly demonstrated that although teachers are indeed important in advancing student learning, there is considerable heterogeneity in terms of their impacts on desired learning outcomes (e.g. Goldhaber and Brewer 1997; Ballou, Sanders and Wright 2004; Kane, Rockoff and Staiger 2005; Rivkin, Hanushek and Kain 2005; Boyd et al. 2006; Podgursky and Springer 2007); this heterogeneity is also idiosyncratic to the usual quantitative measures of teacher "quality." Even though school systems typically distinguish teachers by seniority, experience, certification status, and education level (Hoxby 2002; UNESCO 2004 Vegas and Umansky 2005), it has been clearly shown that these characteristics of teachers are, at best, very weak predictors of a teacher's contribution to student learning (Kane, Rockoff and Staiger 2008; Nordstrum 2015).[2] This appears to hold true no matter where the school system responsible for children's education is located—from Nepal to Nashville, Tennessee. The following sections provide an overview of what we know about how to measure teacher practices and classroom conditions that affect learning outcomes, technical and practical issues related to measurement, as well findings particular to LMICs.

2.2 What we know about teacher practices and classroom conditions that affect learning outcomes

Despite considerable efforts to link teacher characteristics to student outcomes, research has not found a relationship between teachers' background or demographic characteristics to student learning. In their randomized sampling of 65 urban and rural schools in the Lahore District of Punjab, Pakistan, Aslam and Kingdon (2012) found that traditional teacher "quality" characteristics (e.g., certification or training) had no bearing on students' standardized exam scores. On the other hand, "process" variables (i.e., pedagogic classroom practices) substantially affect pupil learning. Kingdon (1996) arrived at much the same conclusion in her earlier analysis of urban state-run schools in northern India: the quantity of teacher education and their formal training or experience had little impact on pupil achievement. These conclusions have also been observed in higher-income countries. Aaronson, Barrow and Sanders (2003), using a large longitudinal dataset that linked student achievement with teacher characteristics in Chicago, Illinois public schools, found that 90 per cent of teacher effects were not explained by any measured characteristics.

2 Recent studies have shown that teaching ability does increase in the first few years of one's career (Rockoff 2004; Clotfelter, Ladd and Vigdor 2006; Jepsen 2005; Rivkin, Hanushek and Kain 2005; Harris and Sass 2007; Aaronson, Barrow and Sanders 2007), but little else in terms of characteristics is predictive of student learning outcomes.

One potential reason that research on teacher characteristics has not provided a better explanation of teachers' contributions to student learning may be that these measures simply do not adequately capture the processes of teaching and learning. If the processes of teaching and learning occur in the interactions that teachers and students have while engaging with academic content, teachers' contributions to student learning might be best measured through structured classroom observations. Understanding how teachers teach and documenting their pedagogical moves in instructional settings would, therefore, be an important contribution to research because although classroom observation has been routinely employed as both a formal and informal means to evaluate and develop a corps of teachers, relatively little is known about how pedagogical moves and instructional behaviour in the classroom relate to student learning outcomes (Jacob and Lefgren 2008; Harris and Sass 2009).

Many studies (discussed in Section 2.2) point to a positive and statistically significant relationship between teacher ratings using observational tools and student achievement data. These studies, while useful as a system monitoring tool, are limited in the sense that the teacher observational data only gives a final score obtained from the observational tools, but cannot comment on the whether certain pedagogical moves, as compared to others, are predictive of student achievement. In other words, the level of granularity is rather high. We know, therefore, that teachers who are rated highly by observers using observational instruments tend to also have students who perform well on achievement tests, but we do not know from these studies which specific instructional techniques, if any, are more effective in enhancing students' understanding of academic content.

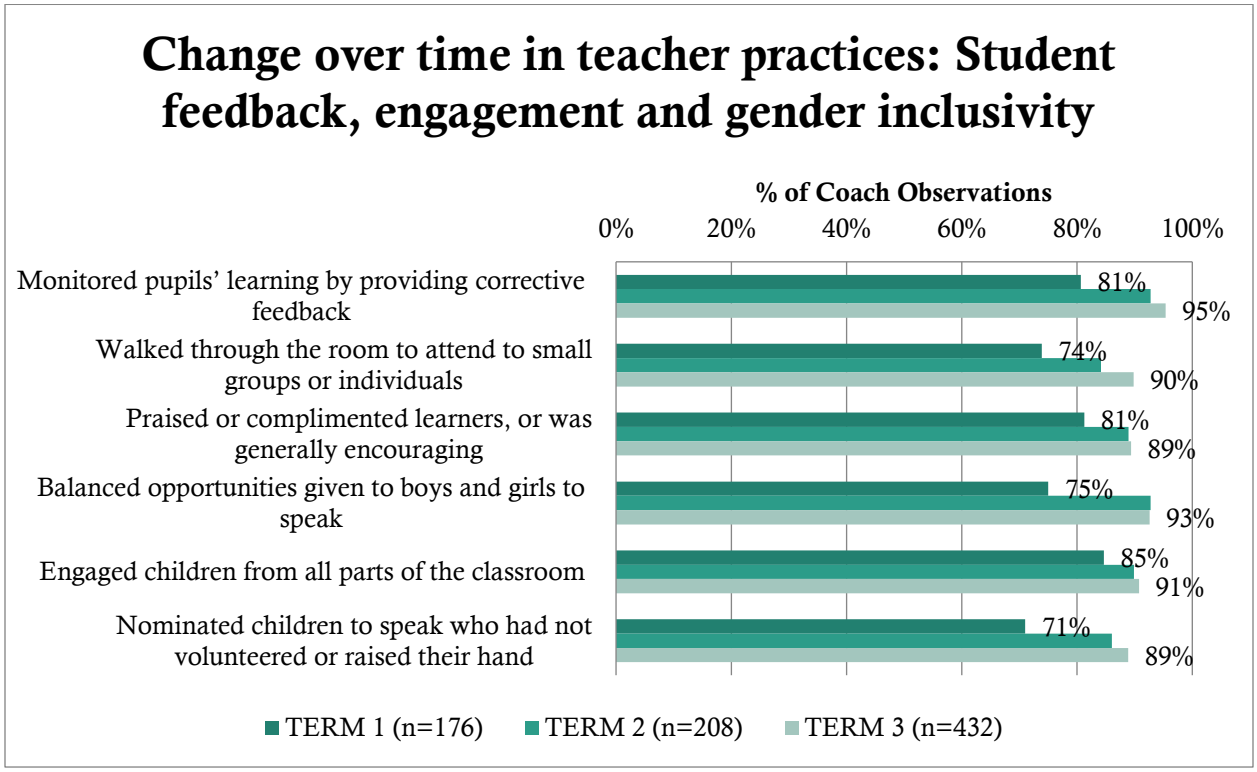
2.1.1 Teacher observational data studies

At least two recent studies conducted by RTI International have attempted to distinguish which instructional practices that teachers employ in the classroom are related to student achievement. One of these studies, the Nigeria Reading and Access Research Activity (RARA) supported by the United States Agency for International Development (USAID) and implemented by RTI during the 2014/2015 school year, tested an approach to improve the early grade reading skills of children in formal government schools in the Northern Nigeria states of Bauchi and Sokoto. The RARA approach was evaluated through a randomized controlled trial: one group of 60 schools (treatment) participated in RARA activities, while a second, similar, group of 60 schools did not (control). Teachers' instructional practices and grade 2 students' reading outcomes were measured at the beginning and end of the 2014/2015 school year, prior to and after implementation of the new reading approach. Assessors observed teachers in both timed and untimed observations. The research found that teachers in treatment schools tended to allocate more time during reading lessons to literacy activities (e.g., 30 minutes of a 45 minute lesson) than their peers in control schools (e.g., 12 minutes in a 45 minute lesson) and had less idle time (5 minutes as compared to 20 minutes in control schools) (RTI International 2016). Teachers' use of effective instructional practices also was assessed, and a score assigned to their implementation of the lesson. s. These "lesson implementation scores" were then analysed vis-vis their students' reading fluency, as measured through an oral reading assessment conducted on the same day as the classroom observation. Students of teachers in the highest lesson implementation score quintile (according to their observational score) were found to be more fluent readers than students with teachers with

scores in the lower quintiles, indicating a clear relationship between the quality of literacy instruction and student outcomes. The effect sizes of the gains in students' oral reading fluency (ORF) scores as a result of teachers' improvement in instruction grew progressively larger from one quintile to the next, with an effect size of 0.38 for the ORF gains among students whose teachers' mean implementation score was in the highest quintile.

As part of the same study in Nigeria, information on teachers' instructional practices as they relate to gender sensitivity and equity, inclusion and child-centred pedagogy were also collected. The results are summarized in ***Figure 1***, which shows the percentage of classroom observations in which teachers were observed conducting or engaging in practices related to providing feedback to pupils, monitoring their work during class, praising or complimenting them, providing equal opportunities to boys and girls to speak in class and engaging pupils throughout the classroom. The data show teacher improved throughout the academic year. Analysing these results vis-vis student outcomes, students whose teachers were observed to be engaging in more equitable, child-centred practices had statistically significantly higher letter sound and oral reading fluency scores, on average, than students whose teachers were not observed using these practices. For girls in particular, those who had a teacher who balanced opportunities for boys and girls to speak had average letter sound identification scores that were more than three times greater than girls whose teachers were observed not to balance opportunities in a gender-equitable manner.

Figure 1: Teachers' Level of Engagement with Students



Note: Observations do not represent unique teachers, but are an aggregate of observations of the same cohort of teachers, which may include multiple observations of the same teacher. (RTI International 2016).

The second study was a 2013 national baseline assessment for the 3Rs (i.e., reading, writing and arithmetic) in Tanzania, a national program that attempts to enhance classroom instruction and student competencies in these content areas. Reading and arithmetic assessments were administered to 2,266 randomly sampled Grade 2 students in 200 randomly sampled schools. Germane to this analysis, the teachers of sampled students were observed teaching arithmetic and reading lessons and their instructional actions were linked to student outcomes. During observed reading and arithmetic lessons, teachers' instructional practices were recorded every three minutes. Brombacher et al. (2014) reported notable differences in terms of the frequency and nature of teachers' pedagogical moves between teachers in high-performing schools (as measured by average student achievement in that school) and their counterparts in low-performing schools. With regard to mathematics lessons, Brombacher et al. (2014) found that teachers in schools associated with high pupil performance also allotted significantly more time to answering pupils' questions and asking questions of pupils.

Although these studies do not make causal inferences into teachers' practices and student achievement, they do attempt to systematically describe differences in instruction that are associated with learning outcomes. In that sense, studies like these begin to fill a sizable gap in the already limited research that links teachers' observed classroom pedagogy with the achievement of their students, either currently or in the future. It is clear that more comparable research would be beneficial to exploring the relationship between specific pedagogical practices that teachers use in their classrooms and student learning. However, classroom observation, as it is routinely conducted in schools and classrooms, is not a panacea; rather, the practice has potential for bias (see below and Section 5).

2.1.2 Studies examining teacher observation and student outcomes

A number of studies have investigated the interplay between student learning outcomes and subjective teacher ratings obtained through classroom observation (Bommer et al. 1995; Heneman, 1986; Gallagher 2004; Kimball et al. 2004; Milanowski 2004). Notably, all of the more recent studies (Gallagher 2004; Kimball et al. 2004; Milanowski 2004) find a positive and significant relationship between teacher "value-added" (i.e., the teacher's unique contribution to student achievement when controlling for demographic information, prior achievement, and [often] peer effects) and student achievement. This result holds true across studies despite variations in the way value-added is calculated and the stakes (i.e., consequences) attached to teachers' ratings.

More recent studies have attempted to use statistical techniques to mitigate possible measurement error in student achievement data. Using a value-added model for measuring teachers' contributions to student achievement, Jacob and Lefgren (2008) combined student-learning data with principals' ratings of 201 teachers in a mid-size school district in the western United States.[3] They found that at both the low and high end of the value-added scale, principals' ratings of teachers matched the student outcomes data relatively well. In the middle of the student achievement distribution, however, principals' ratings of teachers had little to no relation to student achievement (i.e.

3 The district chose to remain anonymous in the present study.

calculated value-added).[4] Interestingly, while on average the relationship between principals' ratings and teachers and value-added was positive and significant, Jacob and Lefgren (2008) also found that teachers' prior value-added scores were better predictors of current-year student outcomes. In other words, teachers' contributions to student outcomes (as measured by value-added) in a current year have a stronger statistical relationship to their contributions to the outcomes of other students in prior years than do principals' ratings through classroom observations. In sum, the addition of principals' subjective ratings added some explanatory power to prior student achievement data in predicting future student achievement.

Harris and Sass (2009) also attempted to compare the ability of current value-added and principal ratings (obtained through classroom observations) to predict future teacher value-added. Their analysis is based on data obtained from interviews with 30 principals (comprising17 elementary schools, 6 middle schools, and 4 high schools) from a diverse midsize school district in Florida, United States. The authors found that when a principal's overall rating of a teacher is added to an explanatory statistical model that includes typical measures of teacher characteristics (e.g., experience, educational attainment, and certification), the coefficient on principal ratings is positive and highly significant in both reading and mathematics content areas, while the coefficients on teacher characteristics remained the same as seen in prior models that did not include principal ratings. While Harris and Sass's sample size was small, their findings suggest that principals may have some knowledge—obtained through observation, but probably also based on prior knowledge of the teacher, their behaviour, and dispositions—about teacher ability that is not captured by standard measures of teacher behaviour in the classroom.

Similar to the aforementioned studies, Rockoff and Speroni (2010) examined the relationship between teachers' observational scores and their students' future achievement using two sets of data to determine whether the positive association between observational data and achievement holds for newly recruited teachers. One set of data on observational evaluations came from the New York City Teaching Fellows, an alternative route to certification that is taken by approximately one in three new teachers in New York City. Teaching Fellow applicants were observed and evaluated using a 5-point scale as part of the interview process. The second source of observational data was a mentoring program for new teachers in New York City. In the program, a trained, full-time mentor met with each teacher on a weekly or biweekly basis in order to improve instructional skills. Mentors submitted both monthly summative evaluations and bimonthly formative evaluations of teachers. Like the Teaching Fellowss, Mentee teachers were also evaluated using a 5-point scale, which was based on a set of teaching standards. The authors found that teachers who received higher subjective scores in observational evaluations conducted before hire or in their first year of teaching were also more likely to have future students with greater average achievement gains. Interestingly, Rockoff and Speroni (2010) also determined that the summative and formative evaluations were highly correlated (coefficient of correlation = 0.84). These findings suggest that mentors and principals can,

4 This finding might be due to the fact that many calculations of teacher value-added cannot statistically distinguish between teachers in the middle of the distribution (Raudenbush and Jean 2012).

on average, distinguish the pedagogical practices of more effective teachers from less effective peers, as measured by student achievement data.

One of the largest and perhaps most well-known studies to investigate the relationship between classroom observational data and other measures of effective teaching is the Measures of Effective Teaching (MET) Project funded by the Bill & Melinda Gates Foundation. The MET Project took place over two school years (2009–2010 and 2010–2011) and across six school districts in the United States (e.g., Charlotte-Mecklenburg in North Carolina; Dallas, Texas; Denver, Colorado; Hillsborough County, Florida; Memphis, Tennessee; and New York City, New York). Before the second year of the project, a subpopulation of students was randomly assigned to teachers who had volunteered for the randomization. The project's overall sample consisted of 1,559 teachers in 284 schools across the six districts (Garrett and Steinberg, 2015). The MET Project conducted an analysis of a subset (1,333) of these teachers who taught mathematics or English in Grades 4–8. These teachers were recorded teaching lessons several times throughout the school year. The recorded lessons were observed outside of the classroom and scored three separate times by different assessors according to structured observation tools. Each lesson was scored using two cross-discipline observation tools ([1] the Framework for Teaching [FFT] and [2] the Classroom Assessment Scoring System [CLASS]) and a third time with the content-specific tool, Mathematical Quality of Instruction or the Protocol for Language Arts Teaching Observations (Measure of Effective Teaching Project 2012).

This analysis yielded three main findings vis-à-vis the relationship between observational data and student learning. First, data from all of the observational instruments were related to student achievement gains in mathematics and English. The project found, for example, that students fell behind their peers by roughly 1 month of schooling in math, all else being equal in classes taught by teachers who scored in the bottom quartile (i.e. below the 25th percentile) in their FFT or CLASS observation scores. Conversely, students of teachers with observation scores in the top quartile (i.e. above the 75th percentile) were more apt to be ahead of comparable students by the equivalent of 1.5 months. However, the correlation between student achievement and teacher observational data was relatively weak—between 0.12 and 0.34, depending on the instrument (Harris 2012). Second, the project found that the differences in teacher "effects," as measured by observational data combined with other measures, on student tests scores were roughly half as large on English tests as in mathematics. Third, the project determined that there was a stronger relationship between classroom observation scores and student scores on open-ended reading assessments than on closed-ended (e.g., multiple choice) reading tests. Although these studies are rigorous and their findings tend to corroborate each other, a limitation inherent in all of them is that the analysis of teacher observational data employs only a final score obtained from the observational tools.

2.3 Experiences and findings from recent research in the global south about teacher practices and classroom conditions that effect learning outcomes

This section summarizes findings from selected studies that have analysed the relationship between classroom observation indicators with learning outcomes, drawing from RTI-managed surveys and datasets. It is important to make the distinction between classroom observation and lesson

observation. Classroom observation, for the purposes of this paper, refers to static aspects of the classroom environment that can usually be observed with simple survey items or checklists, such as availability of books, a print-rich environment, or appropriate infrastructure. Lesson observation refers specifically to procedures that aim to gather metrics related to instructional practice and teacher behaviours, usually through checklists, rubrics or time-sampled lesson observations. The different types of instruments used in LMICs are described in section 3.1.

2.2.1 Program and system monitoring in Malawi

The Malawi Early Grade Reading Activity (EGR Activity, 2013–2016) is a USAID-funded activity designed to provide technical assistance to the Ministry of Education, Science and Technology (MoEST) to improve the reading performance of Malawian learners in Grades 1–3.[5] To accomplish this, the activity works in 11 districts to improve EGR instruction, increase parental and community engagement in providing reading support, and strengthen the policy environment to support early grade reading. Key elements of teacher support include three teacher-training cascades per year, followed by ongoing coaching visits and scripted lesson plans that guide teachers in implementing a phonics-based approach to reading instruction. The coaching visits are conducted by MoEST's Primary Education Advisors (PEAs) who are based at teacher development centres and have responsibility to support instruction in all the schools in their zones, which may contain between 8 and 17 schools.

The lesson observation instrument. Because classroom and lesson observations are a standard part of a PEA's school visit, the EGR Activity and the MoEST developed a teacher observation instrument. The tool directs PEAs' attention to elements of instruction that were considered pertinent to the goal of improving reading instruction. The standard teacher observation tool can be filled out either on hardcopy (i.e. paper) or using Android tablets loaded with an electronic version of the instrument.

Completing the teacher observation is just one part of the process a PEA engages in when visiting a school. For example, prior to observation, the PEA has typically already met with the school's head teacher to brief them on the purpose of the visit. If the visit in question is a follow-up, before the class being observed begins the PEA may meet with the teacher to review particular issues that were noted during the prior visit. Following the observation, the PEA conducts a very brief assessment of the reading performance of up to four sampled students; the outcomes of this assessment serve to nuance and further inform the PEA's discussion with the teacher about instructional practice. The PEA and teacher then meet to discuss the lesson and identify areas of strength and areas for improvement. Once all classroom observations have been completed, the PEA gathers the head teacher and all early grade teachers for a group debriefing where the findings of the visit can be discussed for the benefit of all, rather than just those who were observed. It is important to recognize that classroom observations serve the purpose of instructional support and coaching in a larger ecosystem of support that builds on previous training programs. However, data from these visits can also serve as system monitoring.

5 In Malawi, 'Grades' are known as 'Standards', but for consistency across the paper, we refer to these as Grades.

Observational data. To understand how observational data helps inform program and system design, we looked at data from a total of 4,977 Grades 1, 2, and 3 Malawian teachers who were observed a total of 7,018 times between January 2014 and January 2016. These records were collected and analysed by an RTI statistician using a cross-sectional longitudinal method.[6] The data aims to determine the extent to which teacher behaviour can be affected by coaching visits, under the assumption that teacher behaviour is a key piece of the equation that results in student learning outcomes. However, we cannot, in this instance, associate coaching visits with student outcomes.

The following four indices[7] of observed teacher behaviours were created in order to discover if there was an association between improving observed teacher behaviours and the number of coaching visits they received:[8]

1. Observed teacher preparation
2. Observed teacher reading instruction
3. Observed teacher phonics instruction index
4. General instruction index

Findings. ***Table 1*** shows descriptive statistics for the four indices. Four ordinary least squares (OLS) regression models[9] were created using each index as the independent variable and the frequency of coaching visits as a covariate. Covariates, such as years of teaching experience, teaching qualification, and a teacher's gender, were also tested, but only teacher gender had a statistically significant covariate and was retained in the final model. All the indices demonstrate significant gains observed as the number of coach visits increase.

6 Items which changed significantly across teacher observation instrument versions have been excluded from this discussion. The dataset analyzed for this discussion includes records that were initially filled out on paper and then entered into the electronic system. Therefore it represents all records known to have been generated in Early Grade Reading Assessment (EGRA) schools as of February 2016, regardless of mode of initial collection.

7 An index is created by assigning points to desirable metrics as observed in the classroom or collected through checklists or other data collection metrics. Scores from individual classrooms can then be mapped in terms of how close to the total 'perfect' score they are. For detailed information on index creation, see Varly 2010 (Senegal and Mali Hewlett studies) and Pouezevara et al. 2015 (first Philippines four-country study).

8 These four categories were created using principal component analysis, taking the first eigenvector after removing covariate questions not associated with the others. This index construction is shown in Appendix 1.

9 OLS is a statistical technique that uses sample data to estimate the true population relationship between two variables.

Table 1: Descriptive Statistics of Indices

Index	Mean	SD	Observ-ations	Min	25th percentile	Median	75th percentile	Max
Observed teacher preparation	0	1.23	4,873	-6.73	-0.53	0.19	0.98	0.98
Observed teacher reading instruction	0	1.99	1,527	-2.02	-2.02	-0.22	2.49	2.49
Observed teacher phonics instruction	0	1.69	3,378	-2.13	-2.13	0.86	1.90	1.90
General Instruction Index	0	1.71	4,898	-3.20	-1.45	0.25	1.90	2.29

SD = standard deviation

Table 2 displays the outcomes of the regression analysis. It is notable, that the table's data indicates that female teachers had better behaviours in teaching preparation and general instruction. Effect sizes (Cohen's *d*)[10] were calculated by dividing the coefficient by the standard deviations shown in ***Table 1***. As such, when the number of coaching visits were four or higher, teacher behaviours were almost consistently small or medium effects as compared to a single visit. Although the association between the coaching visit and teacher observed performance is evident, it is more difficult to determine exactly why the teacher performance improved. Two possibilities include good quality coaching and accountability, neither of which precludes the other. However, the table shows improving teacher behaviours as the observation number increases. It is likely that if the teacher was not being coached and accountability was the only cause, the improvement would ceiling quickly.

10 This type of effect size is named Cohen's d, where (typically) an effect size of 0.2, 0.3 and 0.5 are considered small, medium and large effects, respectively (see Cohen 1992).

Table 2: Regression Association of Observed Teacher Behaviour Index and Number of Coach Visits

Response Variable		Teaching Preparation Index			Reading Instruction Index			Phonics Instruction Index			General Instruction Index		
		Coeff	Effect size	n	Coeff	Effect size	n	Coeff	Effect size	n	Coeff	Effect size	n
Female Teachers		0.18°	0.14	—	0.01	0	—	0.06	0.04	—	0.19°	0.11	—
Visit #	**1**	0	—	693	0	—	982	0	—	1,838	0	—	2,673
	2	0.12*	0.10	538	0.31*	0.16	156	0.11	0.07	379	0.34°	0.20	525
	3	0.26°	0.21	166	0.10	0.10	38	0.21	0.12	118	0.41°	0.24	173
	4+	0.32°	0.26	118	0.77*	0.39	23	0.31*	0.19	86	0.50°	0.29	118
Constant		-0.15°	—	—	-0.07	—	—	-0.06	—	—	-0.22°	—	—
F-score		14.15°	—	—	2.33*	p = 0.05	—	2.33	—	—	17.89**	—	—
Obs		4,775	—	—	1,473	—	—	3,301	—	—	4,806	—	—

coeff = coefficient Obs = observations * p<0.05, ** p<0.01, ° p<0.001 n = Number of teachers

Table 3 shows the percentage of teachers observed carrying out a particular activity on a given observation visit for the teaching preparation questions. Although the proportion of teachers that met the criteria for each question generally increased, it was not the case with all questions. For example, a teacher having lesson notes was at its highest in the first visit (64 per cent).

Table 3: Teaching Preparation Per cent Observed by Visit Number

Teaching preparation	Per cent Observed			
	Visit 1	Visit 2	Visit 3	Visit 4 +
Teacher has schemes of work prepared	99%	99%	99%	99%
Teacher has a scripted lesson plan	91%	94%	93%	93%
Teacher has lesson notes	64%	58%	59%	57%
Teacher has assessment records	78%	82%	87%	87%
Teacher has teaching, learning and assessments prepared	81%	84%	85%	90%

Table 4 shows the evolution of teacher behaviours specific to reading instruction. All observations show a general increasing trend with the highest observed percentage of teachers demonstrating the characteristic at a later visit. However, final percentages were lower than desired.

Table 4: Teaching Reading Per cent Observed by Visit Number

Teaching Reading	Per cent Observed			
	Visit 1	Visit 2	Visit 3	Visit 4+
Teacher teaches new words using a relevant strategy (e.g. actions or pictures) to ensure that learners show understanding	53%	57%	55%	62%
Teacher verifies predictions	44%	52%	41%	55%
Teacher asks comprehension questions	54%	55%	51%	63%
Teacher helps learners find answers	52%	57%	54%	69%
Teacher asks learners to predict the story from the title and picture	40%	49%	45%	51%

Tables 5–8 show the individual questions used to create the indices in each section. ***Table 5*** lists the observed phonics teaching characteristics and demonstrates that the only question with little improvement since its first observation was teachers giving pupils an opportunity to practice writing.

Table 5 : Teaching Phonics Per cent Observed by Visit Number

Teaching Phonics	Per cent Observed			
	Visit 1	Visit 2	Visit 3	Visit 4+
Teacher reviews previously learned sounds, syllables and words, adding the new sounds to create words	54%	54%	57%	62%
Teacher is able to manipulate sounds by blending (i.e. putting together), segmenting (i.e. taking apart) using onsets and rimes (adding/removing initial letter to the stem)	54%	57%	57%	64%
Teacher demonstrates on the chalk board the mechanics of how to write letters	43%	50%	54%	61%
Teacher provides opportunity for learners to practice writing/drawing, e.g. in the air, on the ground or in students' notebooks or slates	57%	60%	59%	60%

Table 6 shows general instructional teacher behaviours across coaching visits. Although improvements were noted and most teacher instructional characteristics were observed most frequently at later visits, the initial percentages and subsequent gains were mostly modest. However, the results clearly indicate areas of focus for future visits, along with consistent successes, such as teachers following the *I do/We do/You do* procedure.

Table 6: General Instruction Per cent Observed by Visit Number

	Per cent Observed			
General Instruction	Visit 1	Visit 2	Visit 3	Visit 4+
Teacher uses the lesson cycle, e.g. introduction, an advance organizer and the I do/We do/You do procedure for each activity and conclusion	70%	78%	80%	80%
Teacher varies class organization (e.g. group work, pair work, and individuals sharing work) to maximize learning	52%	58%	62%	64%

Teacher supervises and supports learners through immediate and appropriate feedback	65%	71%	69%	71%
Teacher uses teaching, learning and assessment resources effectively	51%	54%	55%	55%
Teacher uses appropriate and gender sensitive language	72%	74%	75%	69%
Teacher assigns appropriate class exercise and/or homework	45%	53%	58%	71%
Teacher uses appropriate pace to cater to learners with different learning and special needs	51%	58%	58%	62%

2.2.2 Survey results to inform system monitoring in Tanzania [11]

The National Baseline Assessment for 3Rs in Tanzania, led by RTI through funding from USAID, has already been introduced. Below we provide examples of how observation data can be compared against student learning outcomes to inform system monitoring through baseline and follow-on data collections.

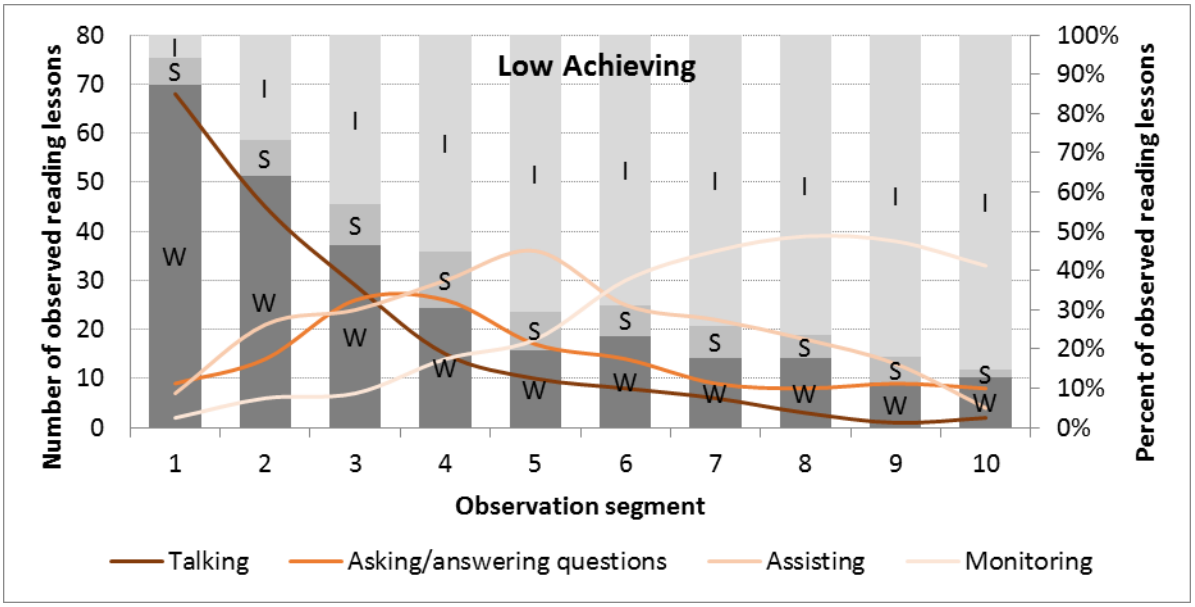
Teaching during observed reading and mathematics lessons. During observed reading and mathematics lessons, teachers' instructional practices were recorded every three minutes. Although the observational categories included greater nuance, teacher instructional behaviours fell into four overarching categories: (1) talking, (2) asking or answering questions, (3) assisting pupils and (4) monitoring pupils. Instructional grouping (e.g. whole group, small group or individual work) was also noted. The classroom observations paint a picture of the general pedagogic approaches employed by teachers over the course of a 30-minute lesson in schools associated with either low or high academic performance. Notable differences were found between teachers in low- and high-performing schools.

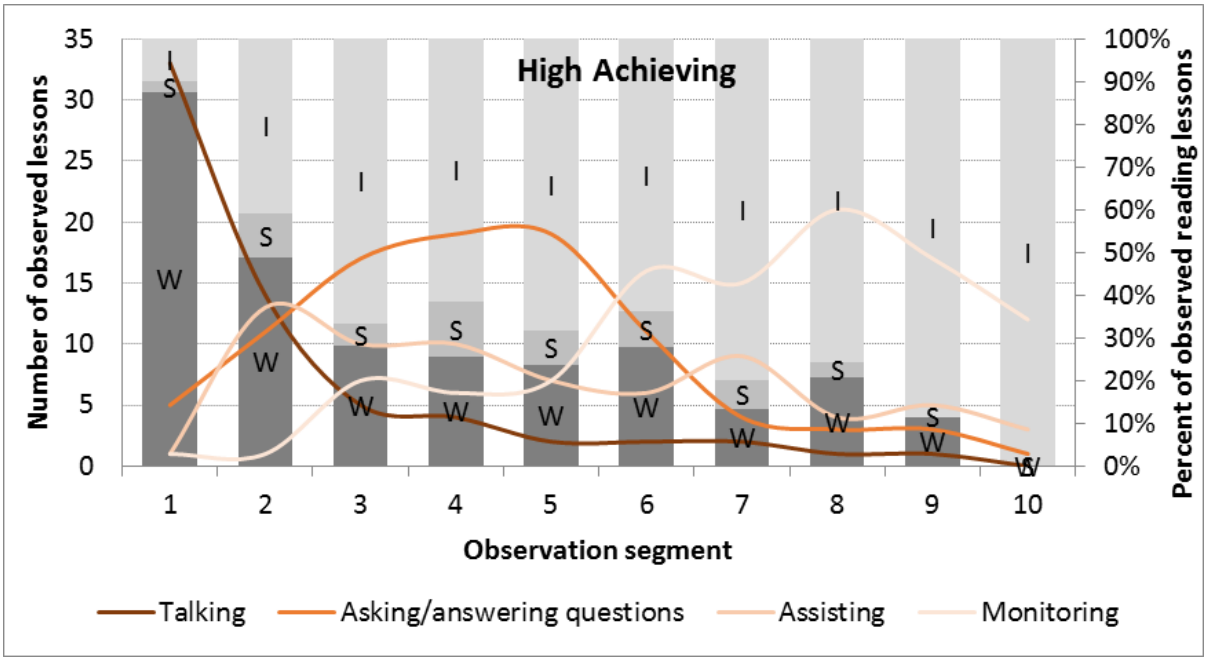
Figure 2 provides observational results for low- and high-performing schools, respectively. The line graph depicts evolutions in teacher instructional behaviour over the course of reading lessons and is aligned to the primary (left) vertical axis. The proportional bar graph represents the instructional grouping used during the observed lesson and is aligned to the secondary (right) vertical axis. Reading lessons in schools associated with both low and high performance tend to commence with the teacher talking or presenting the lesson content in a didactic manner to the entire class. Over the course of the lesson, the teacher transitioned to individual work, during which the teacher was assisting or monitoring individual pupil work. The instructional grouping observations supports these general statements: at the beginning of lessons, most classes are working as a whole group, though the vast majority then transition to individual work over the course of the lesson. Very few

11 This section is an excerpt from the previously published report by Brombacher et al. 2014.

classes, either in low- or high-performing schools, employed small groups during observed reading lessons.

Figure 2. Instructional Practices During Observed Reading Lessons in Low and High Achieving Schools





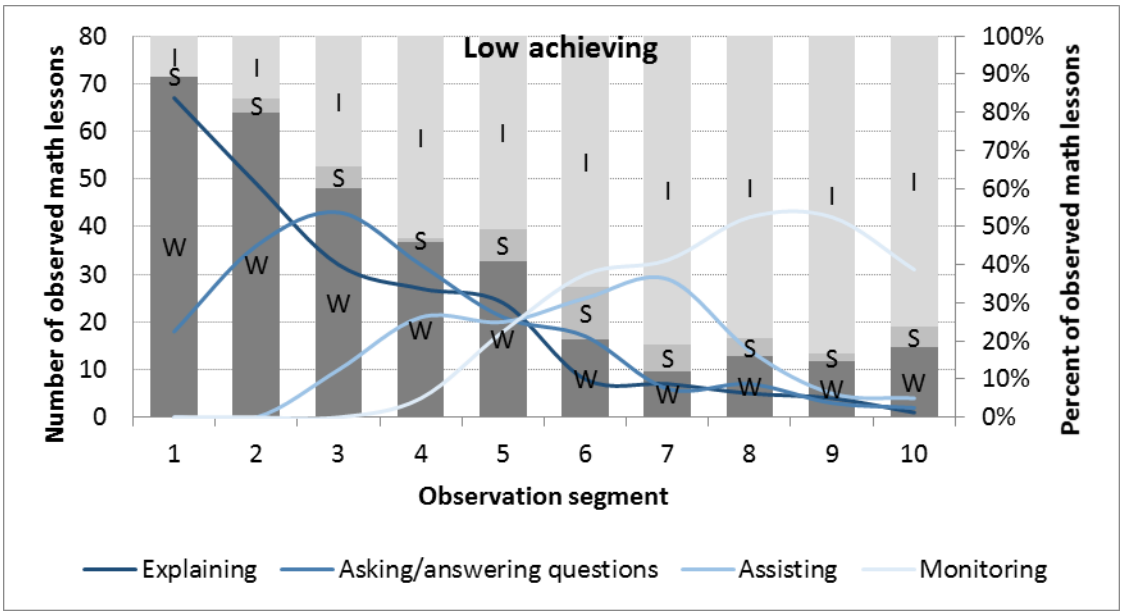
Note: W = whole group, S = small group, I = individual. Line graph associated with primary vertical axis. Instructional grouping (whole group, small group, individual) associated with secondary vertical axis. Number of classrooms observed in high-achieving schools = 85 (Brombacher et al. 2014).

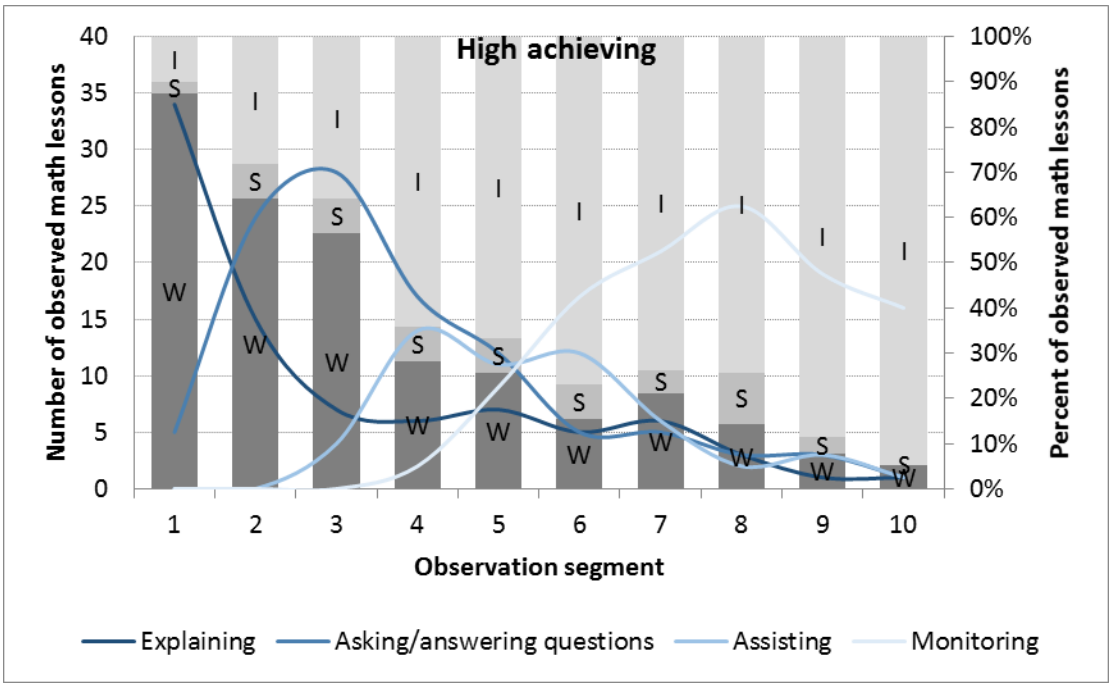
Differences are evident, however, between classes in schools associated with low performance and those associated with high performance. For example, while most teachers in both low- and high-performing schools began lessons by direct instruction, more observed teachers in high-performing schools were able to transition much more quickly to other instructional approaches that allowed for more student engagement (e.g. questions). In classrooms of low-performing schools, this transition did not take place until approximately 10 minutes into the lesson, while in high-performing schools it occurred within six minutes. Classroom teachers in high-achieving schools also tended to allot more time to answering pupils' questions and posing some of their own during the middle of the lesson. Further, observed teachers in high-performing schools were more likely to interrupt the flow of the lesson to address pupils' concerns or questions regarding the assigned task or lesson content (observed in 55 per cent of classrooms); this was less likely to occur with teachers in lower performing schools (observed in only 33 per cent of their classrooms).

The results of observed mathematics lessons for classroom teachers in both low- and high-performing schools are in ***Figure 3***. Similar to the observed reading lessons, mathematics teachers tended to begin mathematics lessons with whole group direct instruction and transitioned over the course of the lesson to question and answer periods followed by individual work time. The transition to individual practice and task work took place at the same juncture in low- and high-performing schools, approximately halfway through the 30-minute observation period. Teachers rarely utilized small groups in observed mathematics classrooms; the vast majority of lessons consisted of whole group instruction followed by individual practice. Unlike reading lessons, however, mathematics teachers very rarely observed interrupting lessons to address pupils' questions or misconceptions: less than one in ten mathematics teachers were observed doing so. Instead, they tended to monitor and assist individual students.

Several findings from the classroom observations of mathematics lessons diverged by school performance. First, teachers in schools associated with high pupil performance allotted significantly more time to answering pupils' questions and asking questions of pupils. Indeed, asking and answering questions was the most frequently observed teacher behaviour in four of ten observation segments (i.e. 40 per cent of the time) in high-performing schools' classrooms, which was double the proportion in classrooms in lower-performing schools (i.e. 20 per cent). Second, teachers of classrooms in high-performing schools transitioned more quickly from one section of the lesson to another than did teachers in low-performing schools, ostensibly maximizing instructional time. Further, teachers in lower-performing schools were observed spending more time explaining mathematics concepts to the whole group at the beginning of the lesson. Third, pupils in classrooms of high-performing schools were provided more time for individual practice and execution of the assigned tasks, with assistance and monitoring, than their counterparts in lower-performing schools. This was, in part, due to the efficient transitions between lesson segments in classrooms in high-performing schools.

Figure 3: Instructional Practices During Observed Mathematics Lessons by Low and High Achieving Schools





Note: Note: W = whole group, S = small group, I = individual. Line graph associated with primary vertical axis. Instructional grouping (whole group, small group, individual) associated with secondary vertical axis. Number of classrooms observed in high-achieving schools = 85; classrooms observed in low-achieving schools = 40 (Brombacher et al. 2014).

Table 7: Factors Associated with EGRA-EGMA Performance tabulates factors from the child-, classroom-, and school-level survey instruments that were found to have a statistically significant association with high Early Grade Reading Assessment (EGRA) and Early Grade Mathematics Assessment (EGMA) performance. This relationship was tested via a logistic regression model that, in essence, assesses what affected a student's achievement on the EGRA and EGMA. The "factor" column in ***Table 7*** lists determinants that were found to have a statistically significant relationship with a student's either high or low performance on EGRA and EGMA. The "per cent of sample" column presents the proportion of the survey populations that either reported having or had (depending on the survey) the factors in column one. The final two columns indicate, according to our models, how much a factor's presence increases the likelihood of whether the pupil is a high- or low-performer on both EGRA and EGMA. For example, the first factor, "living in an urban location", can increase the likelihood that a pupil will be identified as a top EGRA-EGMA performer. Put another way, pupils in tested urban locations were 4.8 times more likely to be top performers on the assessments conducted.

Table 7: Factors Associated with EGRA-EGMA Performance

Factor*	Sub-category	% of sample	Increased likelihood of pupil being an EGRA-EGMA:	
			"High performer"	"Low performer"
Urban location	-	26.6	4.8	—
Household wealth index	Low	25.0	—	3.0
	Med-low	25.3	—	2.3
	Med-high	24.7	10.0	—
	High	25.0	15.1	—
Pupil reads aloud at home	Never	32.9	—	7.6
	Every day	19.3	7.7	—
Pupil is read to at home	Never	35.9	—	4.6
	Every day	12.3	3.3	—
Pupil attended pre-school		80.3	2.6	—
Pupil brings textbook to class	Kiswahili	11.1	2.4	—
	Math	11.9	2.5	—

High parental involvement		52.2	2.8	—
School performance classification	Low	87.2	—	4.5
	High	1.4	10.1	—
Teachers trained		28.9	3.5	—
Teacher absentee rate	>15%	44.0	—	1.7
School year begins with appropriate materials		10.9	3.8	—
% of exercise book pages marked by teacher	None	3.9	—	8.3
	25%	36.4	—	4.7
	50%	19.0	—	3.9
Teacher gives positive feedback to pupils		51.4	2.6	—
Teacher checks for pupil understanding	Math	76.4	10.2	—
Pupils participate actively in class	Reading	12.1	3.9	—
High % of pupils respond to questions in class	Reading	0.5	10.2	—
	Math	6.8	4.9	—
Advanced pupil discussion in class	Math	0.2	27.8	—

*Relationships tested by logistic regression models (i.e. Does the factor increase the likelihood that pupils are high- or low-performing?).
All relationships reported here are statistically significant at the 0.05 level.
Grey rows = factors that are related to classroom observation.

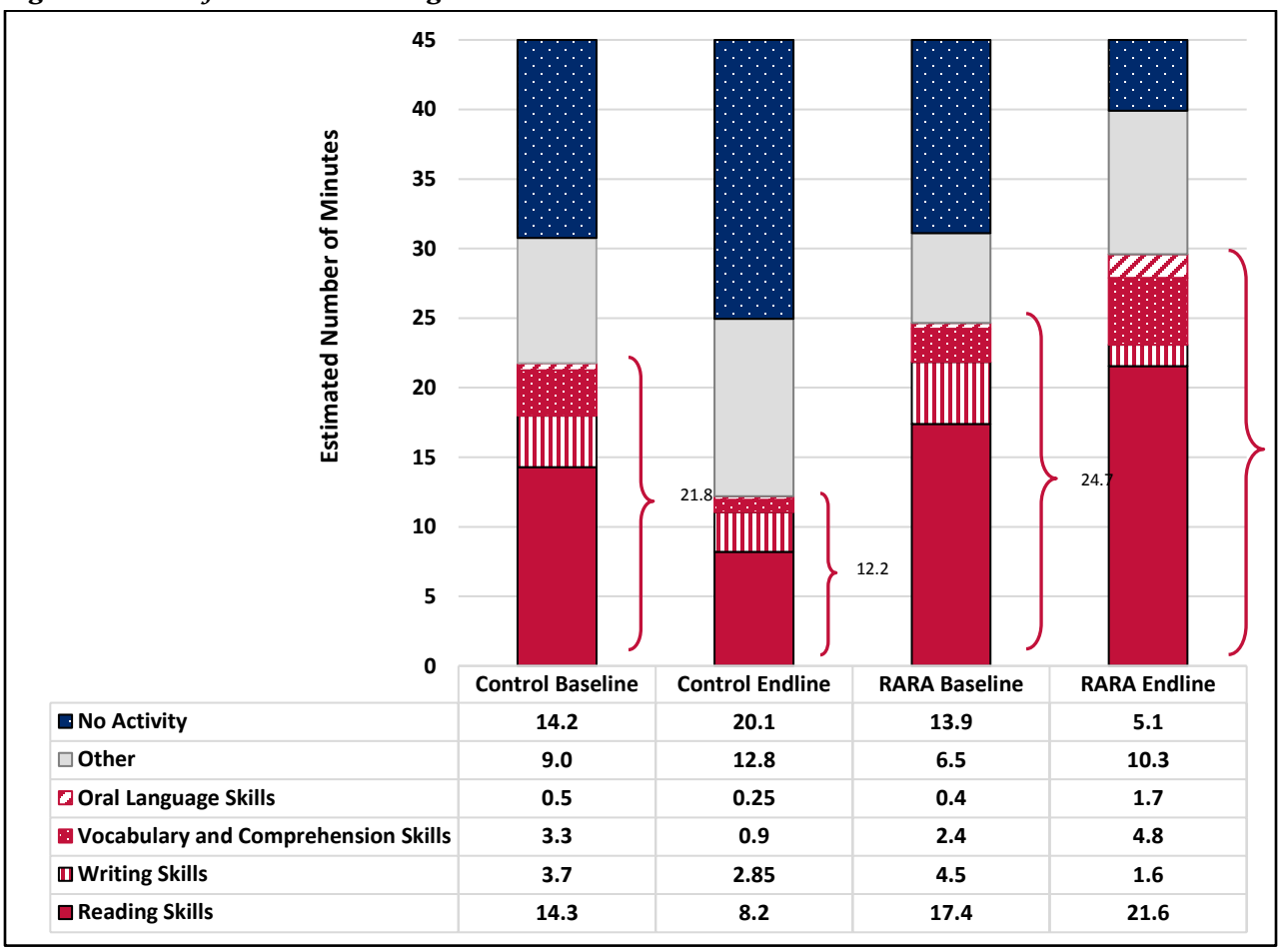
As shown in ***Table 7***, numerous factors had an association with an increased likelihood of either high or low performance on EGRA and EGMA. Those factors that were related to classroom observation appear in grey, of which only the last four of the six were collected through lesson observation. The observational factors with highest correlation to positive student performance were all related to the process of active student engagement through questions, response and discussion. Pupils whose teachers checked whether they understood the content taught during mathematics lessons were more likely to be among the highest-performing students. In addition, pupils' active participation in class (i.e. those that engage not just when called upon, generally answer questions correctly, pupil-pupil engagement and putting forth arguments and defending them) were associated with an increased likelihood of high pupil performance.

2.2.3 Program monitoring and evaluation in Nigeria

As part of ongoing teacher support and monitoring provided through the Nigeria RARA initiative, reading coaches periodically observed teachers' delivery of an early grade reading lesson and recorded what they saw on a teacher observation tool. This information recorded was used "on the spot" by the reading coaches to provide feedback to teachers immediately after the lesson. It was also later analysed to determine whether teachers were implementing the lessons as planned (a measure of fidelity of implementation) and whether their instruction was improving. This aggregated information was relayed back to both teachers and coaches to inform them about progress and gaps. Areas where teachers seemed to be lagging behind were targeted for additional training during a mid-academic year workshop, as well as for targeted support from coaches. Because reading coaches only indicated whether these practices were observed during a lesson, the data do not necessarily indicate the *extent* of a teacher's behaviour or the *quality* of a given reading instruction practice. Nevertheless, the questionnaire items served as a prompt for the reading coach to take notice of issues and, ideally, provide feedback to teachers. More detailed or nuanced items would be required if the main purpose of the classroom observation in Nigeria were to assess the quality and extent of teacher practices in these areas.

As discussed earlier, (see Section 2.1.1) the Nigeria RARA initiative also used teacher observation to evaluate the impact of the initiative in improving student achievement on EGRA. On the same day that teachers were observed, their students' reading skills were measured through the administration of EGRA. Subsequent analysis indicated that in classrooms where teachers used a greater number of effective instructional practices, student outcomes tended to be higher. Specifically, for every one-point increase in a teacher's lesson implementation score (a score based on the number of effective instructional practices, out of 12, that the teacher was observed using during a given lesson), a student's letter sound identification score increased concurrently by an average of 0.51 letters per minute. This is equivalent to an average effect-size increase of 0.06 for every extra additional effective instructional practice (as measured by the observational instrument). Student oral reading fluency scores were also higher when teachers used a greater number of effective instructional practices. These findings indicate that data collected via a classroom observation survey instrument can be a reliable predictor of student reading gains. Given this relationship, such a tool could be used to indicate whether a teacher's instruction is likely to improve students' reading skills. It would not, however, necessarily indicate students' actual reading scores, but it would provide a way to gather some information as to whether a teacher's instruction is "on track" to improve student reading outcomes.

Figure 4. Focus of Instruction During Timed Classroom Observation



	Control Baseline	Control Endline	RARA Baseline	RARA Endline
No Activity	14.2	20.1	13.9	5.1
Other	9.0	12.8	6.5	10.3
Oral Language Skills	0.5	0.25	0.4	1.7
Vocabulary and Comprehension Skills	3.3	0.9	2.4	4.8
Writing Skills	3.7	2.85	4.5	1.6
Reading Skills	14.3	8.2	17.4	21.6

Note: The number of minutes spent on each activity is projected from the data recorded at 15 intervals every 3 minutes for 45 minutes.

***Error! Reference source not found.*Figure 4** shows how classroom observation data can be used to evaluate and quantify how much time during a given lesson teachers use to teach particular content or skills. Data presented in the graph were collected from the Nigeria RARA timed classroom observation instrument and illustrate change over time in the focus and duration of instruction between the treatment group receiving training and materials on reading instruction and a control group. The red bars indicate time devoted to literacy instruction, the grey bars indicate a non-literacy activity was taking place, and the blue bars indicate no activity was taking place (e.g., the lesson ended). This particular way of presenting the classroom observation data allows for the actual amount of time to be quantified and compared across time, groups and specific reading and writing skills taught. For example, the data show that treatment school teachers devoted, on average, more than twice as much time to literacy instruction than teachers in control schools (29.7 minutes compared to 12.2 minutes).

2.2.4 Observation of language of instruction

Another example of an adaptation of a time-sampling instrument (similar to that described in Section 2.2.2) for use in LMICs was the Monitoring Learning Outcomes in Developing Countries project implemented by RTI for the Hewlett Foundation. In four countries—Mali, Senegal, Uganda,

and Kenya—researchers measured language of instruction in the classroom. Data were collected every two minutes and data capture only focused on what was happening precisely at that time interval.[12] The snapshot format provided a picture of overall language use and the extent to which code-switching was used by teachers even in the context of one 'official' language of instruction mandated by the school system. Due to low variation in language use, no relationship was found between language use and EGRA (French) scores in Senegal. Nonetheless, other types of findings that were generated through this combination of instruments included:

- Teachers used national languages more during math lessons than in other subjects,
- With few exceptions, teachers used Wolof in the classroom only when the proportion of Wolof speakers exceeded 50 per cent,
- The use of French was not related to the proportion of pupils speaking French at home or the diversity of the linguistic situation in the classroom, and
- Classroom instructional time was largely lecture-based, with teachers speaking to children who were simultaneously writing (e.g. copying off the board onto their slate).

Select variables from the classroom observation tools and student and teacher questionnaires were used to create an "opportunity to read" index as detailed below.

- One person at the student's home could read (student reported)
- Pupil had reading textbook (observed) and pupil had other books at home (student reported)
- Teacher gave homework (student reported)
- Time spent reading per week at school (teacher-reported)
- Teacher absence (teacher-reported)
- Pupil absence (student-reported)
- Literacy supports on the walls (e.g. posters, decorations, and/or pupils' writing)

This index was incorporated into regression models 'in order to compare teachers' method size effects with pupils' literacy background (including availability of reading materials and support)' (Varly 2010, p. 21). The analysis compared the extent to which EGRA scores on various subtasks could be explained by certain factors, including:

12 Data does not record what teachers and pupils were doing during the last two minutes, but only at the specific point in time. Therefore a limitation of this method for capturing language use is that often, brief code switching was done between the 'snapshot' moments, and was not recorded. Data collectors made notes to indicate that this happened, but there was no way to estimate the exact proportion of time that this might have happened.

- student characteristics (pre-schooling and repetition),
- opportunity to read,
- teacher and school characteristics,
- classroom effects (using a dummy variable), and
- teachers methods and language use

Overall, most of the factors in Senegal that were significantly associated with reading score outcomes were not those that were collected through the classroom observations (although time spent writing was positively and significantly correlated). Instead, factors, such as teacher and student absenteeism, repetition and teachers' expectations explained most of the variance.

2.2.5 Monitoring the implementation of mother tongue-based multilingual education in the Philippines

In 2013, the Philippines Department of Education launched an ambitious series of reforms, which included the expansion of basic education from 10 to 12 years and the introduction of mother-tongue-based multilingual education curriculum (MTB-MLE). In the MTB-MLE model, children begin school with mother tongue as the medium of instruction and are gradually introduced to Filipino and English before shifting completely to Filipino and English as the medium of instruction in Grade 4. As part of these reforms, new curricula, textbooks, teaching guides, and classroom resources were rolled out to schools. Under a partnership between USAID and the Department of Education, RTI used a combination of EGRA and lesson observations to help inform implementation of the new approach. EGRA data were collected from regionally representative samples in four languages.

The lesson observations were conducted in Grade 1 and 2 classrooms on the same day as EGRA administration; observed teachers were those whose students were sampled for the reading assessment. Additionally, teacher and student interviews were carried out to gather other contextual data (such as home language, presence and use of books at home and in the classroom) to associate with the observations and reading assessments. An 'implementation index' score was derived from lesson observations and contextual data allowing the researchers to categorize schools on a scale, or degree of reform, along three components: (1) teacher training, language ability and the support teachers receive at their schools; (2) materials available and being used in classrooms and (3) instructional practices observed during reading lessons. For example, the metrics gathered through classroom and lesson observation make up a 26-point scale for the latter and include whether

- mother tongue is in use for 80 per cent or more of the class time;
- pupils are actively engaged in the lesson;
- the teacher models proper mother tongue use for pupils to emulate, praises or rewards pupils for using mother tongue correctly, points out and corrects mother tongue errors, monitors pupil comprehension and uses visual aids and simpler forms of the language to explain

something the students struggle to understand (the index would award a point for each positive behaviour observed);

- the teacher shifts to another language only in order to clarify or emphasize a concept first explained in mother tongue or to give procedural directions to students;

- Twenty-five per cent or more of a reading lesson is spent on activities that involve students reading and

- during a reading lesson, students also spend time productively listening, speaking or writing in their mother tongue.

To account for possible differences between languages, RTI identified the top 25 per cent of students based on oral reading fluency within each language. We then used logistic analysis to determine how home- and school-related factors affected the likelihood of a student being in this subset of readers. We found a statistically significant relationship between the combined index score (component 1, 2 and 3, combined) of a child's environment and their being in the top 25 per cent of readers in the region, even when adjusted for demographics, such as gender and age (RTI International 2015). On the other hand, individual components, or items within the components, did not yield significant results, due largely to sample size.

During two years of data collection using the same tools, this approach was able to provide the Department of Education with information on how the new MTB-MLE policy was being rolled out, i.e. whether teachers were adopting and implementing the new requirements, if instructional practice was shifting due to reforms and where gaps remained.[13]

3. Types of lESSON observation instruments that have been applied to assess classroom teaching practices in LMICs

13 See also RTI International 2016.

Globally, there is a multitude of classroom data collection protocols and techniques that differ in terms of their relative emphasis (e.g. on students, teachers, interactions, environments or time allocation), question type and format for coding responses. For example, simple checklists can be used to record the presence or absence of clearly observable behaviours or classroom inputs (e.g. books, desks and writing materials). On the other hand, time-sequence lesson observation tools typically specify not only the presence or absence of a behaviour, but also capture the frequency of occurrence. Systematic classroom observation, increasingly used in developing economies to monitor classroom practice involves coding observable teacher or student actions over a defined period within a lesson. The purpose of observational studies may be descriptive, investigative or evaluative and, as such, can involve low- or high-stakes. The purposes may sometimes overlap and the data can potentially be used in many different ways, as described below.

> "Systematic classroom observation is a quantitative method of measuring classroom behaviors from direct observations that specifies both the events or behaviors that are to be observed and how they are to be recorded. Generally, the data that is collected from this procedure focuses on the frequency with which specific behaviors or types of behavior occurred in the classroom and measures their duration." (Waxman, et al., 2015)

- **Training or other types of situation analysis**: Trainers observe current practices to determine training needs or application of training concepts post-training (example of descriptive, low-stakes use). Researchers may conduct observations in order to better understand current practices so that they can help design an intervention, training curriculum or other program inputs (example of investigative, low-stakes use).

- **Pedagogical coaching**: Observations seek to identify strengths and weaknesses with respect to teachers' instruction with the intention of stimulating reflection and dialogue between the teacher and coach on how to improve classroom practices (example of descriptive, low-stakes use).

- **Monitoring**: A formal means by which observers record the extent of teachers' adherence to a particular curriculum or program. When observation is conducted for monitoring purposes, a set of pedagogical and/or other classroom practices is observed and associated with effects on students; observation determines whether classroom characteristics align with known effective practices (example of investigative/evaluative, low-stakes use).

- **Evaluation**: The act of appraising whether teachers' practices or classroom environments meet the quality requirements of a particular standard, which may be used to make personnel decisions or resource allocations; determining whether a particular intervention,

curriculum or program results in positive changes (example of evaluative, high-stakes use).

Two approaches to data collection may be used to apply these observations to monitoring the quality of education, depending on the purpose and context.

- **Surveys**: Sample-based, system-level data collection activities, usually conducted by professional enumerators external to the education system. The purpose is typically to inform an external evaluation and is often conducted along with student learning assessments or other sets of questionnaires.

- **Monitoring**: Embedded approaches, usually conducted by personnel from within an education system or project, such as district supervisors or instructional coaches on a routine basis.

Although the same instruments can be used for multiple purposes, experience suggests that careful alignment with local capacity—that is, the capacity to observe subtle differences in subject behaviours, capacity to capture data (e.g. on paper or electronically) and capacity to manage the data that comes out of these instruments—is as important as ensuring the right fit for purpose. Differences in content and format will reflect both the purpose of the assessment and the wider context in which it is being used.

Two key differences in question format concern the level of inference, or subjective judgement, that an observer makes when recording the observation data. *Low-inference* tools are used to gather information on processes and actions between the teacher and students without drawing immediate conclusions. Instruments focus on directly observable actions and can use easy-to-code formats, such as checklists and binary response (i.e., yes/no or true/false) observation items, or more detailed written records that objectively document everything that happens. These can be used, for example, to document the fidelity with which teachers implement something they are expected or taught to do (Al Otaiba et al. 2006, p. 230). On the other hand, *high-inference* tools are typically more informal, subjective and open-ended (Al Otaiba et al. 2006, p. 230). The assessor is expected to make a judgment about the occurrence and quality of the indicators and in place of binary-response items will use likert scales for coding or other scaled global ratings like 'unsatisfactory' through 'exceptional'. The level of inference refers primarily to the content of the question, not its format; even a binary-response question can be high or low inference. For example, 'Teacher behaviour supported a positive classroom environment' is high inference, whereas 'Teacher smiled and had encouraging words at least four times' is low-inference, but both would be answered with 'yes' or 'no'. Low inference questions and time-sampling of discrete behaviours tend to have higher reliability than global ratings or other high-inference measures (Stuhlman et al. n.d., p. 5). Geddes (2015) suggests that global ratings scales should include at least 5 to 7 points in order to increase reliability and differentiate performance adequately. Some protocols may require a combination of a low-inference 'running record' containing detailed observations to support the high-inference ratings. The validity, reliability and feasibility of different options are the types of considerations that go into

the selection or development of observation tools and will be reviewed in more detail in Section 3.2.2.

In this section, we review some of the types of instruments that have been used in developing economies and how they demonstrate alignment to these specificities. In Section 5, we summarize recommendations for their effective classroom observations.

3.1 Sample classroom observation tools

Table 8 summarizes descriptions of some of the most common observation protocols that are used to monitor classroom quality in developing economies. This list is not exhaustive, but represents a range of formats with documented results. Following the table, we discuss five of the protocols in more depth.

Table 8: Summary of classroom and lesson observation tools used in LMICs

Standardized tools used across projects							
	Question type		**Data input type**				**Other**
Name[14]	**High inference**	**Low inference**	**Time-sample**	**Global ratings**	**Checklist**	**Running record**	**Embedded feedback**
SCOPE	x			x (5-point likert)			
TIPPS	x			x (4-point, double-dichotomous)		x	
Stallings		x	x (intervals)				
MELQO		x	x (cycles)				
SSME		x	x (intervals)		x		
Tangerine: Tutor		x			x		x
Simple Interactions	x			x			

14 Definitions for all abbreviations are found in the protocol descriptions following the table.

Project-specific tool adaptations					
		Data entry format		Results reported	
	Based on tool type	Paper	Electronic	System level	School level
Nigeria RARA classroom observation	Stallings		x	x	
Nigeria RARA checklist	SSME				
Hewlett four-country language analysis	Stallings	x		x	
Kenya TAC Tutor classroom monitoring	Tangerine:Tutor		x	x	
Malawi EGRA	Tangerine:Tutor	x	x	x	
Tanzania 3Rs baseline	SSME		x	x	
Philippines 3-year study	SSME		x	x	

Standards Based Classroom Observation Protocol for Educators in Literacy (SCOPE Literacy). Developed by Educational Development Center (EDC), the SCOPE tool was first trialed during a USAID-funded program in Egypt with a classroom instruction focus of a more generic nature, not focused on literacy. SCOPE-Literacy was then developed to specifically address literacy instruction, and has been used in the Philippines. It is a classroom observation tool that assesses teaching practices in language and literacy instruction (EDC 2013). SCOPE Literacy has evaluated the impact of project and ministry activities on teachers' instructional practices through both summative and formative evaluation exercises and supports instructional supervisors to support continuous professional development by identifying gaps in teacher behaviours in two main areas of practice: (1) Classroom structure (6 items) and (2) language and literacy instruction (7 items). SCOPE characterises itself as a complex, high-inference instrument that records information about teachers' observable behaviours. Observers rate items on a scale of 1 (deficient) to 5 (exemplary) based on defined criteria and indictors.

Teacher Instructional Practices and Processes System (TIPPS). Developed by Dr Edward Seidman (New York University) and teams participating in the Equitable Access to Quality Basic Education (OPEQ) intervention trial in the Democratic Republic of Congo and Global TIES (Transforming Intervention Effectiveness and Scale) projects in Uganda, Ghana, and Pakistan, TIPPS is an observation tool to measure classroom environments (e.g. teacher practices and classroom process) across subject areas. It can be used for impact evaluations; research studies and to provide teachers feedback for teacher training, support and supervision. TIPPS includes nineteen key concepts that indicate the quality of (1) emotional/social support in the classroom environment, (2) teacher-provided opportunities for learning and personalization of the material, (3) facilitation of students' cognitive development, (4) equitable treatment of students, and (5) student engagement (New York University 2016). Like SCOPE, TIPPS is a high-inference tool that uses a 4-point ratings scale with descriptive indicators and examples to scaffold observer coding.

Figure 5 is an example of one category of observation in the TIPPS instrument.

Figure 5: TIPPS Categories and Examples

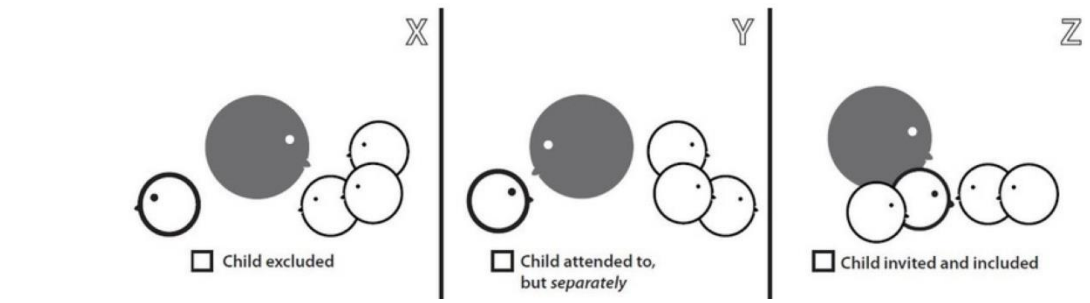
Category A Examples	Category B Examples	Category C Examples	Category D Examples
The teacher does not incorporate pupil interests or suggestions into the lesson.	The teacher, at times, may allow pupils to express interests or suggestions about the lesson but lesson may still highly teacher driven.	The teacher encourages pupils to express and discuss their interests or suggestions, yet, incorporation into the lesson is only intermittent.	Teacher regularly allows for pupil expression and tailors the lesson in a manner that incorporates pupil interest.
In a lesson about farm animals, the teacher lists a variety of animals (cow, goat, chicken, etc.) in a lecture format.	In a lesson about farm animals, the teacher says, "Raise your hand if you have a chicken at home." Pupils raise their hands, but the teacher then moves on with the lesson.	In a lesson about farm animals, a pupil asks, "We have a chicken. Is that a farm animal?" Teacher responds, "Yes, that is also a farm animal..." The teacher then moves on with the lecture.	In a lesson about farm animals, a pupil asks, "We have a chicken. Is that a farm animal?" Teacher responds, "Very good. That is also a farm animal. And can you tell us what you could do with a chicken on a farm?"

To address the low reliability of scaled responses, TIPPS is designed with a double-dichotomous rating format that reduces subjectivity. The two-step decision-making process can be discerned by the example above, where Columns A and B are (broadly) negative ("no") responses to the first question of interest, while Columns C and D are affirmative ("yes") responses. Each column then provides a second level of determination of the degree of accuracy in each category (e.g. 'somewhat' in columns A and C and 'very' in columns B and D). The protocol encourages observers to take notes in addition to using the coding sheet to have a 'play-by-play' of the observation period. The data identifies areas where a majority of teachers are struggling or performing well, which in turn helps governments design interventions that can address those gaps and acknowledge best practices. TIPPS is also used during instructional coaching to provide immediate feedback to teachers. Factor analysis has identified four variables associated with a latent variable (i.e. quality basic education): (1) instructional support, (2) learning connections, (3) structures for deeper learning and (4) socioemotionally supportive environment.

Stallings-inspired tools. The Stallings 'snapshot' model of classroom observation was developed by Jane Stallings in the 1970s to measure use of classroom time in the US (Stallings 1977). In 2002, the World Bank began to adapt the tool for use in developing country contexts to answer educational questions raised in education projects (Texas A&M University 2001). The Stallings snapshot focuses on student, teacher and classroom level interactions and classifies each activity (and the intended audience of that activity). Data is captured at regular intervals (usually every 2-5 minutes) wherein interactions are recorded as though through a photograph—a single instant in time is the focus of the observation. This makes it a low-inference, quantitative tool for measuring time on task. However, the number of items to score can increase the complexity of the instrument and reduce reliability.

Another limitation of the time-on-task focus means it does not necessarily provide information about the quality of that time on task, but it can be a useful proxy for quality to the extent that such studies can draw attention to time *off* task. For example, ***Figure 6*** shows time-sequence data gathered from a classroom observation tool used to measure the impact of a an early grade literacy improvement initiative in Northern Nigeria (RTI International 2016). As discussed in Section 2.2.3 the classroom observation tool was used to capture information about how time was used by teachers in treatment and control schools, before and after a reading improvement initiative was implemented.

Figure 6. Example of Simple Interactions tool



Source: Akiva and Li 2016

The Stallings format has been adapted by RTI for use in other contexts and for other objects of study and is commonly associated with the School Snapshot of Management Effectiveness (SSME) process, including teacher interviews and classroom visits. Although the format remains the same across countries, the observational focus can change. For example, the focus can vary from measuring the breakdown of instructional time in early reading classrooms between practice with reading, writing, listening and speaking, as in the Philippines in 2014 and 2015 (See section 2.2.5) to determining the percentage of teacher's attention directed towards girls versus boys or individuals versus groups, as in Jordan in 2015 or Morocco in 2014, respectively.

Simple Interactions. This methodology[15] is designed as an approach to identify positive interactions between adults and students in a variety of educational settings. The strengths-based approach emphasizes four key characteristics of 'positive' developmental interactions:

1. **Connection** – interacting with mutually positive or appropriate emotions
2. **Reciprocity** – balancing roles of engagement during joint activity
3. **Progression** – presenting incremental challenges and matching with appropriate support (scaffolding)

15 Developed by the University of Pittsburgh (School of Education and Office of Child Development) in collaboration with Fred Rogers Center for Early Learning and Children's Media and Allegheny Partners for out-of-school time. It consists of a tool to be used with a school-based program of professional development. www.simpleinteractions.org

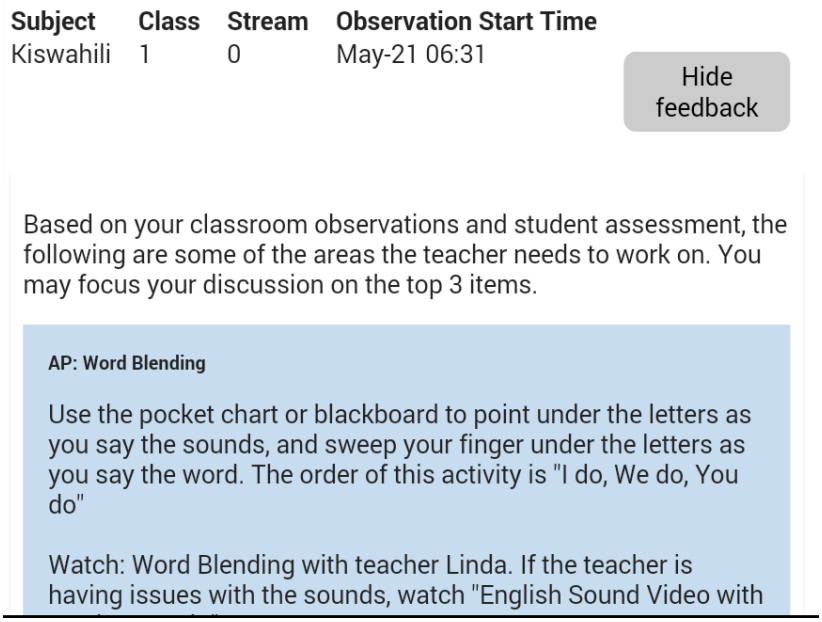
4. **Participation** – inviting and involving all students regardless of ability, temperament or other factors

The observation tool is designed as a very simple ratings scale that is depicted through pictograms with little text so that it does not rely on literacy skills in order to be useful. Methodologically, the Simple Interactions process recommends using video recordings to capture classroom interactions and subsequently selecting positive interactions to discuss at the school level in peer group workshops. The tool has, to date, been used in early learning centres in China, but has not otherwise been used widely in LMICs. It is not designed as a system-level tool (wherein data from separate observations is aggregated to report and generalize on the state of practice in a country or region). Because of the emphasis on simplicity of the observation instrument it may be more easily adopted across different contexts, languages and levels of experience.

Tangerine®:Tutor. In LMICs, RTI frequently uses observation protocols for system monitoring purposes that measure multiple dimensions of classroom environments, including teacher behaviors, teacher-student interaction, and overall classroom organization and infrastructure. These types of observational protocols are often used in the US, and include. CLASS [Pianta and Hamre 2009] and the Framework for Teaching Observation Survey [Danielson 1996]). Data from the tools can be analysed by item or aggregated into an overall score. In one study of over 3000 US classrooms (PK-5), students taught by teachers with higher CLASS scores made greater academic and social gains (Pianta and Hamre 2009). Therefore, monitoring the classroom environment can be an important continuous diagnostic with which to plan appropriate interventions. Yet, in order to do so, observers, teachers and others data users need to know how to interpret the individual items and scores and how to effectively respond to findings.

To address this challenge of turning data into appropriate feedback and interventions, RTI converted a custom lesson observation checklist to digital format for mobile data capture designed to support visiting instructional coaches ("tutors") in monitoring classroom environments for reading instruction. An extension of the Tangerine® survey software, Tangerine:Tutor's functionality goes a step beyond the structured observation protocol by also presenting the observer with automatic suggestions of feedback to give to the teacher based on observations inputs. Tangerine:Tutor has thus far been adapted for Kenya and Malawi. In Kenya, it is used nationally by all government instructional coaches, known as Teacher Advisory Centre (TAC) tutors, for both continuing professional development and accountability purposes. The tool is primarily designed to be used to observe whether teachers are adhering to the expected reading instruction lesson script through 'yes/no' questions (e.g., 'Did the teacher instruct the children to read from the book?'). However, it also gathers information about certain choices made during instruction (e.g. 'What type of activities were modelled?') and elements of the classroom environment (e.g. 'Do all children have a book?'). The TAC tutor observes the reading lesson and enters data directly into a tablet computer loaded with the Tangerine software. At the end of the observation, to help guide the coaching feedback session, the software automatically recalls for the TAC tutor key elements of the lesson that were skipped or altered. ***Figure 7*** shows a sample feedback screen.

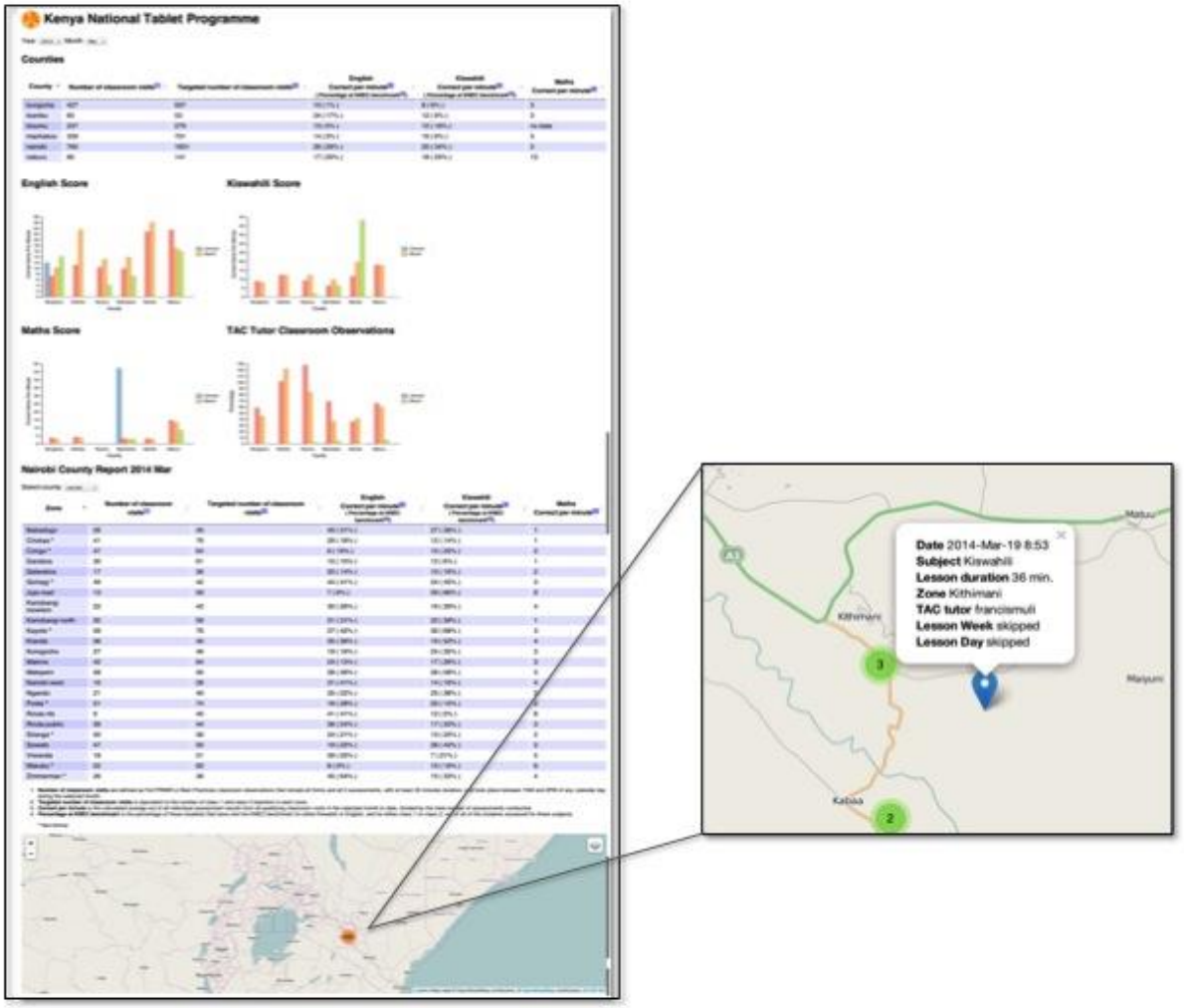
Figure 7: Sample Feedback Screen from Tangerine:Tutor used in Kenya



Although the observation tool could easily be filled out on paper, the technology has the added value of being able to process and display the feedback, as well as direct the coach to specific supplementary resources available on the data collection device, such as model videos that can be used to provide further support to the teacher during the feedback session. The data is not aggregated into a score; instead, the process of data gathering serves as an aid to structure teacher feedback based on minimum performance expectations. Moreover, tutors can upload data from the observations to a central server, which serves as a system accountability tool. In addition to the observation data, the tablets collect global positioning system (GPS) data and school name, zone and region information, which can be aggregated and reviewed to give policymakers an overall picture of system-level support.

Figure 8 shows a screen shot of Tangerine: Tutor's online reporting, including results for one calendar month. As shown in the figure, results are aggregated by zone and country with visualizations, including comparison of visits over a 3- month period. Further, a dynamic GPS map allows location-based monitoring, providing details on the latest synced data for the location.

Figure 8: Screenshot of Classroom Observation Visits in Kenya



Measuring early learning and quality outcomes (MELQO). The MELQO assessment tool is under development by a consortium led by the United Nations Children's Fund, UNESCO, World Bank and the Brookings Institution with partnership on technical development from Save the Children, the Inter-American Development Bank and independent experts representing universities and organizations from many regions. When finalized, MELQO will be a validated tool that measures children's learning and development, provides feasible, actionable data on learning and development at the start of formal schooling for children between the ages of four and seven years and provides information on the quality of learning environments in the years immediately preceding enrolment in formal schooling (i.e., pre-primary or pre-school). It has been piloted in Tanzania, Bangladesh, Mongolia, Lao PDR, Sudan and Madagascar. At the time of this publication the format, question types and scoring methods were still in development, but the basic construct is a combination of low-inference observational questions that gather information about the presence or absence of empirically derived indicators of developmentally appropriate early childhood classroom inputs and processes across subject areas. Although most questions are not time sequenced, but apply to the whole observation period—usually a full day or half day of instruction—some questions are designed to be 'cyclical' so that they are answered more than once, by subject area. For example, specific questions will pertain to the reading, math, art or science lessons in the class and a question, such as 'Teacher smiled, clapped and had warm words of praise for children's efforts', will be answered for each subject lesson to provide a measure of frequency of some interactions. The data from the observation tool can be aggregated into a score, or quantitative 'index', of the classroom environment that can be used to inform professional development and monitor improvement over time.

3.2 Considerations and recommendations for classroom observation development and use

3.2.1 Strengths

Classroom observations, either alone or in combination with other types of results-focused measurement tools (i.e. student-level assessments), provide important insights into the broader classroom ecosystem that ultimately impacts the quality of education and its outputs. This is because the classroom and teacher are necessarily unavoidable *mediators* in the translation of direct inputs, training programs or policy directives and reformers must have access to what is inside a classrooms in order to effect change. For example, the presence of textbooks or new technologies do not, on their own, influence student learning outcomes. Rather, it is how the inputs are used—through teacher mediation—that makes a difference. Similarly, teacher training programs are not automatically applied in the classroom without the *mediating effect* of existing classroom and school structures and even students. Classroom observation provides an opportunity to explore the processes of teaching and learning in action. If carefully constructed, classroom observation tools can also be a practical way to support teachers through feedback on their practice. Moreover, careful construction of the instruments can serve to articulate clear expectations for performance improvement for teachers and school leaders, thereby developing a shared vision and vocabulary with which to promote effective teaching (Hamre n.d.).

To ensure strengths are maximized, several issues need to be considered when identifying the appropriate content and design of a classroom observation instrument.

- **Purpose of the data collection**. Data on how teachers are using instructional time can be used in many different ways. As such, the content and design of an instrument are inextricably linked to what information is desired and how the data will be used. This is important for both time and resource considerations, as well as to guard against unnecessary complexity (for administrators of the tool and data analysis).
- **Instrument development process.** To ensure that classroom observation instruments are developed with both the purpose and the end users in mind, the instrument development process should include people who are expected to use the instrument to gain an initial sense as to whether the amount of content, the vocabulary and language used and the overall design/layout of the instrument is appropriate. Further, field testing or piloting of a draft instrument is recommended to gain insight into its content and use. Data gathered during the instrument pilot can also be analysed to identify whether the instrument is capturing the required information. Finally, a pilot also provides opportunity to assess the data collectors' capacity and helps target appropriate training.
- **Frequency and time allocated for observation**. The number of times that a classroom observation needs to be conducted and the amount of time that data collectors will need to collect it will both influence the content and features of the instrument, as well as the selection and training of data collectors. If teachers are to be observed multiple times during the year in order to provide them with support and to track progress, instrument content and features should be designed to help coaches provide immediate, actionable feedback to teachers. The amount of time a data collector has to observe a given lesson will also be an important factor to consider in terms of the amount of information that can be collected. This may be compounded by the frequency of observations, as well as instrument length and complexity. Therefore, consideration should be given to both how much time data collectors have to observe a given lesson and how frequently they will be expected to observe.
- **Skill level of data collectors**. The skill level of those who will be conducting the observations and collecting data is another important consideration, as agreement among enumerators evaluating responses and recording data, known as inter-rater reliability, is generally lower for the administration of a classroom observation instrument than, for example, a reading assessment, due to both the nature of the content and the scoring procedures for the instrument.[16] For example, the reading coaches who observed teachers' instruction under Nigeria RARA (Sections 2.1.1 and 2.2.3) had little background in reading pedagogy, classroom observation and data collection. As such, the content of the instruments needed to strike a balance between providing robust and useful data, while at the same time being simple enough for data collectors to capture reliable and valid data. In contexts where the

16 See Brown et al. 2010.

skill level of data collectors is fairly low, it is recommended that instruments be relatively short with minimal complexity so that enumerators can be adequately trained and to reduce the likelihood of error (which in turn improves data accuracy and reliability).

- **Data management and analysis**. The amount of time needed to manage and analyse data, and the availability of skilled personnel to do this, should factor into the design of a classroom observation instrument, the frequency of data collection and how results will be used. As instrument complexity and length increases, so too will the time and skill requirements for analysing the data collected, and for communicating results. Data that cannot be (or is not) analysed and used represents wasted time, effort and money.

3.2.2 Limitations: Technical and practical considerations

Technical issues using classroom observation as a means of measuring teaching practice

When determining whether to use a given measure to collect data on teacher practices, we should be clear as to the purpose for which the measure is to be used and informed by the technical properties of the measure. While technical properties (e.g. reliability and validity) are important, not all purposes require equivalent technical specifications. For example, using lesson observations to evaluate teachers or schools (i.e. making personnel decisions of consequence) requires a higher degree of reliability and validity than does using observations for formative and informal instructional coaching. Decisions of the former type (formal evaluations) become problematic if they are taken without high levels of reliability and validity because classification errors—classifying ineffective teachers as effective and vice versa—can and do occur (Harris 2012; Harris 2013).

It is intuitively reasonable that measures used to assess teaching practices should be, on average, reliable; we should expect teachers who are observed and judged to exhibit more ambitious pedagogical practices at one point in time to exhibit ambitious pedagogical practices at another point in time. In addition, it is desirable that a teacher is scored similarly by two different observers, particularly if those observers are observing the same lesson and employing the same observational tool. However, observational measures of teaching practice are, in practice, not reliable. The reliability of single observations has been calculated to range from a correlation coefficient lower than 0.2 when observations are conducted by trained external assessors to 0.65 when conducted by the teacher's principal (Harris 2012; MET Project 2012; Whitehearst, Chingos and Lindquist 2014). To achieve a similar level of reliability with trained external observers, data from four years of observations would have to be aggregated together (MET Project 2012). Observations from principals and Head Teachers are probably more stable from year to year because principals (and assessors) use some level of prior knowledge of the teacher (e.g. habits and dispositions) to arrive at their final determination (Whitehearst, Chingos and Lindquist 2014).

Validity (or the degree to which a measure gathers data on what it purports to) is also an issue in classroom observation. Bias might affect observational measures of teaching practice in a number of ways. First, as noted above, observations are routinely conducted by staff members who know the teacher well, such as principals or coaches. While this background knowledge of the teacher in question might contextualize the observer's rating of the lesson and pedagogical practices, it may

also bias the rating by incorporating factors unrelated to good teaching (e.g. whether the observer likes the teacher) (Harris and Sass 2009). There is some evidence that this is indeed the case in certain situations. Varma and Stroh (2001) found that certain classes of workers (e.g. women and older workers) tend to receive lower subjective ratings that appear unrelated to their actual ability. Further, using data from the MET Project, Ho and Kane (2013) determined that school administrators tended to score their own teachers more highly than teachers from other schools (0.33–0.66 standard deviations higher). Second, bias may arise from the difficulty inherent in making observational judgments across qualitatively different classroom contexts. For example, a principal or Head Teacher likely has to observe teachers who face dissimilar classroom challenges (e.g. more challenging classroom management situations are likely to result in lower observational ratings, all else being equal). Using data from four US school districts, Whitehearst et al. (2014) found that teachers of students with higher prior achievement levels received higher subjective ratings than teachers of students with lower achievement in previous years. This ratings differential cannot be attributed to the teacher but rather to the non-random sorting of students into schools and classrooms that is common in school systems. Additionally, observational measures might be biased by the relationship between the observer and the teacher. As mentioned above, Ho and Kane (2013) showed that a teacher's direct supervisor rate teachers more favourably on observational measures than do principals from other schools, but even these ratings were also significantly higher than ratings given by peers.

Given the complex nature of classroom interactions between teachers and learners, it is perhaps unsurprising that the act of capturing data on these interactions is also complex. Yet it is sobering that the tools in all of the studies cited in this section are frequently used, well known and previously validated classroom observation instruments. Nevertheless, there are ways to mitigate, or at least limit, the technical issues presented here. Reliability can be enhanced by using more than one observation and aggregating the teachers' score across several observations or by combining observational ratings with other measures of effective teaching. Validity is increased by using external (and trained) observers who do not have substantial prior knowledge of the teacher. However, it is clear that observational measures of teaching practice are subject to numerous forms of bias. This knowledge should inform any attempt to employ observations for the purposes of monitoring, coaching or evaluating.

Practical issues using classroom observation as a means of measuring teaching practice

Several practical and conceptual issues related to the use of classroom observational measures as a means of monitoring, coaching or evaluating teachers are worth mentioning here. These often impinge upon the systematic use of classroom observation to measure teaching practice, particularly in LMICs. First, cost is self-evidently a constraint: high-quality observational measures require extensive training of assessors, many hours spent collecting and analysing data and must be conducted by individuals present in the classroom (as opposed to questionnaires completed by school staff). In sum, they are heavily resourced and time intensive—factors that should be considered alongside the technical ones noted above. Although web-based virtual coaching is becoming common in some high-resource contexts like the US, it may be a long time before such methods are feasible or cost effective in LMICs.

Second, while it may be desirable from an external validity perspective to find correlations between measures of teaching practice and measures of student achievement, high correlations might not be desirable from a practical perspective. After all, if observational ratings are routinely highly correlated with student learning outcomes, it becomes difficult to justify the collection of both measures, particularly given the aforementioned costs constraints (Harris and Sass 2009). A pertinent consideration, in this regard, is how much additional information the observational measures provide, the nature of that information and the cost to obtain it.

Third, there is a conceptual rationale for the undesirability of high correlations between observational measures of teaching practice and student learning outcomes. Intuitively, these measures capture information from different aspects of the teaching and learning enterprise and while it is reasonable to assume that student learning and classroom instruction are related, they are not the same. It has been shown that it is commonplace, in both educational research and policy circles, to conflate effective teaching with successful teachers (Nordstrum 2015; Cuban 2013). The latter (successful teaching) entails that teachers are successful at facilitating their students' progress toward agreed-upon learning outcomes while the latter (effective teaching) involves, at least, teacher characteristics and school inputs, teacher professionalism, student learning and teaching practices. Any robust measurement framework that seeks to provide information on teaching practice should incorporate this distinction (Nordstrum 2015).

Fourth, managing and making sense of data can be a challenge. How data is used, who has access to it, whether they have the capacity to interpret the data correctly and whether they risk making erroneous inferences should be key concerns. High-inference tools in particular create an opportunity for analysts to make inferences that go beyond the scope of the tool. Addressing issues of data management and privacy may require significant capacity building in situations where classroom observation as a monitoring tool is expected to be embedded in the system long-term.

4. Applications and results of classroom observation instruments to directly monitor teacher classroom practices at a large scale in LMICs

4.1 Instrument features

The previous sections introduced common purposes and formats of classroom observation tools to monitor the quality of learning in developing countries, as well as several examples where such tools are currently used. This section provides more specific guidance on the selection of a monitoring mechanism of teacher classroom practices, including practical implementation considerations.

As described in Section 3, identifying and developing the right instrument format, content and implementation model depends on the purpose of the classroom observation, but ultimately it is also dependent on the local implementation context, including capacity and resources. The following considerations should guide development of a suitable tool or combination of tools for each context, whether used by governments or other implementing organizations in the context of periodic surveys or ongoing monitoring.

Reliability. If classroom observation data will be collected and aggregated for reporting, then implementing organizations should ensure that instruments demonstrate high reliability across observers and over time (unless change is expected). Stuhlman et al. describe reliability in observational assessments of classrooms in the following way: 'a tool that produces reliable scores will output the same score regardless of variation in the classroom that is outside of the scope of the tool and regardless of who is making the ratings'(Stuhlman et al. n.d.). This means that the tool must measure 'stable' constructs that are relevant throughout the school week or school year regardless of the particular classroom situation on the day of the observation. A tool that accurately represents teachers' behaviour rather than a random occurrence or a more subjective interpretation will be more likely to enable actionable feedback and support that resonates with teachers and helps them change their practice (Stuhlman et al. n.d.). The degree of subjective interpretation among observers can be reduced with low-inference tools, clear and comprehensive explanations, intuitive data collection formats and sufficient training. There should be clear and standardized protocols for setting up the observation, recording observations and eventually scoring or tabulating global results. Observer training needs to include opportunities to demonstrate that different observers will assign the same rating or otherwise mark a given item similarly when faced with a similar observed behaviour. There may be some scenarios where the classroom observation tool serves primarily as a cognitive aid for the observer, to scaffold the process of providing feedback to the teacher. In this case, reliability between observers is not as important as consistent scoring by each observer—global ratings scales can be more appropriate to clearly articulate performance expectations for both observer and teacher.

Relevance (validity). There is a lot that goes on in classrooms and not all of it will be relevant to the purpose of the observation. Instruments should be carefully aligned to the purpose (which may include specific research questions in some cases) so that outputs or scores derived from the tool relate in a meaningful way to outcomes of interest. If the purpose of observation, for example, is to monitor whether teacher behaviour promotes effective student learning, the observed behaviours

need to have been shown to be related to student learning (i.e. the outcome of interest).[17] A tool designed for that purpose might not be valid for measuring whether teacher behaviour supports gender sensitivity or reduces dropout rates (different outcomes of interest). It is particularly critical that there is strong prior evidence of validity when observation data is used for accountability or evaluation.

Sensitivity. A particular aspect of validity, but more specific to the use of classroom observation for impact evaluation, or measuring change over time, is how sensitive it will be in detecting change due to an intervention. Would change be determined by an upward shift in the rating scale or overall fewer instances of unobserved positive behaviours? How big of a shift indicates positive change? Are changes more likely to demonstrate shifts in a specific teachers' practice or between different teachers? For example, Stuhlman et al., noted that time-sampling methods (Stallings-type tools) are more sensitive to differences in a given teacher's practices at different times, while global rating methodologies are more likely to differentiate between teachers based on stable observed characteristics.

Feasibility. Even after validity and reliability concerns have been addressed, the ideal instrument may not be practical to use in every context. Feasibility of the tool refers to the practicality of filling out the observation tool in the classroom, as well as the logistical considerations that go into getting the observers to the classroom in the first place. Among the key questions in establishing feasibility are: How much time is required to complete the tool on a given day? With what frequency does the tool need to be administered in order to provide reliable information? What are the opportunity costs of the time or resources spent getting to and remaining in the classroom? What are the impacts on classroom practice of having an observer in the classroom and can observers use the tool reliably at scale? Carefully aligning the tool to the purpose and capacity of the assessors to establish reliability and validity will hopefully also ensure that you have reduced the scope of the tool to the minimum necessary observation items for the purpose at hand. Feasibility of a particular tool can be compromised by 'scope creep' or when a tool attempts to do too many things at once.

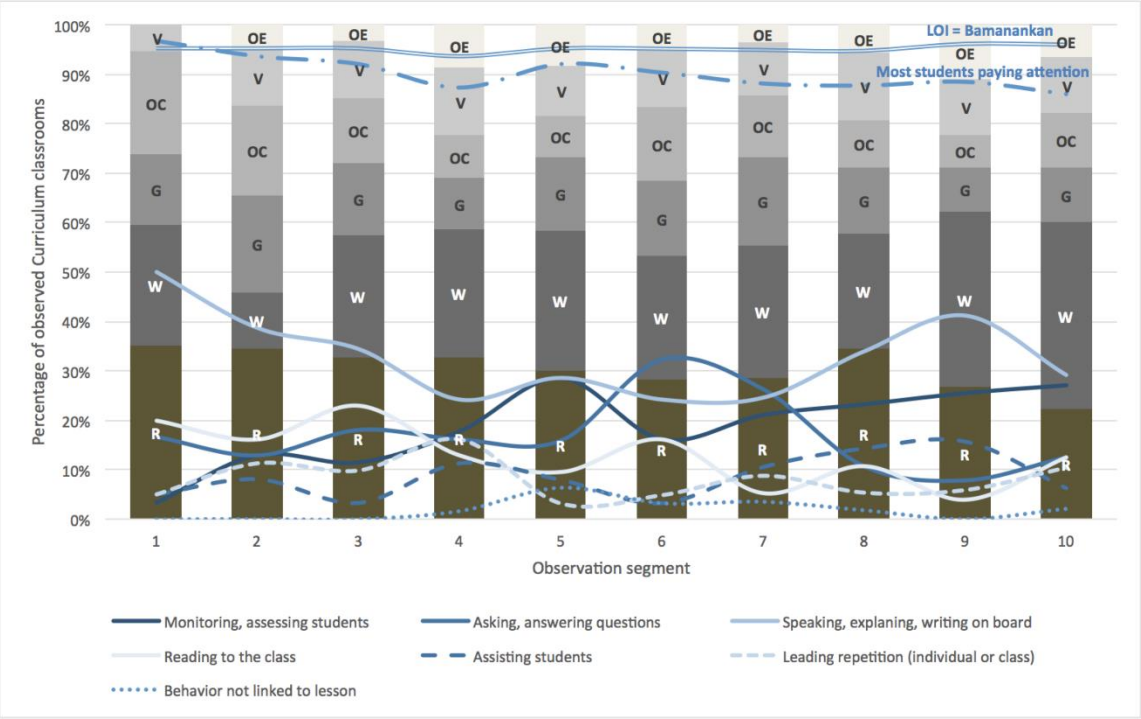
Usability. Reliability and feasibility are both affected by the 'user-friendliness' of the tool; that is, how easy it is to fill out the tool accurately under potentially difficult circumstances. Usability in this context refers specifically to usefulness of the outputs from the tool both practically and theoretically. First, the data generated during classroom observation may be recorded on paper or in an electronic data collection system (RTI International, 2016). If the observation is recorded on paper, the response items need to be encoded in a database for analysis and then tabulated or scored. Data entry may be relatively straightforward for closed, quantitative questions, but for more detailed, open-ended observations or running records of classroom interactions this can be more time consuming. Electronic data collection may ease the data entry burden, but someone needs to be trained to extract the data from the device and process it into meaningful tabulations. This second aspect of usability is even more complex for open-ended questions where deriving accurate and objective meaning be even more complex; it is therefore not practical at scale, although it can be extremely useful for research studies or other infrequent, sample-based observations. Time-sampling data is also more complex to analyse than standard checklist or survey data since it can be tabulated

17 Classroom observation can certainly also be used in a research setting to determine those links.

by observation segment, by variable of interest or as a depiction of the whole lesson sequence. The complexity of analysis increases when aggregating across classrooms and over time.

Figure 9 shows an example of how data was tabulated using a lesson observation tool that measured various aspects of reading instruction in Mali in a representative sample of schools.

Figure 1: Output from a Classroom Observation Tool in Mali



Note: R = reading; W = writing; G = grammar; OC = oral comprehension; V = vocabulary; OE = oral expression.

Source: Pouezevara et al. 2015

The data in ***Figure 9*** provides a robust picture of classroom instruction in this context, but it may be less apparent how to act on the data from this observation. Inferences certainly cannot be made from the chart without understanding the context, curriculum and expectations, among other things. Therefore, the target audience for the data must be determined in advance so that an analysis and reporting plan is aligned to the audience's capacity to extract meaning from the output generated and act upon it to improve instructional practice. Some classroom observation tools can be designed to help evaluators provide high-quality feedback to their teachers either through inbuilt decision trees (see the example of Tangerine:Tutor in Section 3.1) or by articulating precise and explicit expectations (see next paragraph on credibility).

Credibility. Finally, the best way to ensure that the content, format and output of classroom observation tools are aligned to the context and target audience is to involve the stakeholders (e.g. teachers who benefit from the observations) in development and piloting of the tools. Doing so will help ensure credibility (also known as 'face validity') of the tool and uptake of the recommendations that are generated by the data. Moreover, the process of developing the tools can be an exercise in articulating clearly the standards and performance expectations they are supposed to measure. The

New Teacher Project (2011) emphasizes that classroom observation tools should help 'focus teachers and evaluators on the most essential characteristics of excellent teaching'; in other words, 'what teachers need to be able to do to boost the learning of all students,' and not just the minimum characteristics of compliance to a certain program or policy.

5. Conclusion: Priorities and next steps for tools to observe teacher classroom practices

Classroom and lesson observations are increasingly used in LMICs to support improved quality of education by providing information about current practices and measuring change in practices over time as a result of interventions such as teacher training, policy developments or other direct inputs. Data from observational protocols can be used to improve training programs, design effective interventions, support instructional coaching, monitor implementation of reform, measure change in practice or hold teachers and schools accountable for delivery of quality education. The actual costs of data collection through classroom or lesson observations, while potentially high, needs to be thought of in terms of the relative cost-effectiveness of the observation process in providing high quality information on the desired processes, which may ultimately improve teaching and learning outcomes more rapidly.

Classroom observation instruments are generally quantitative in nature, as in a checklist or global ratings scale, but can include qualitative notes in the form of a running record of everything that happened during the observation period. Some tools measure the frequency of specific activities or behaviours and data is collected at specific periods of time in either a timed snapshot (i.e., Stallings) methodology or cycle of observations (i.e., MELQO). Observation tools may gather information on broad, universal standards of behaviour or more subject-specific ones, but in general the format of the tool is not subject dependent; rather, the choice between a timed-bound or cyclical observation, a global rating or simple checklist format is more a function of the research question, the feasibility of implementation and the local capacity for quality data collection and use. In particular, the use of high-inference versus low-inference data collection methods or question types will depend on the level of experience of the observers and the how well they can be reliably trained. The timing and frequency of observations will also depend on the purpose and resources available. Given the growth in the use of observational protocols in LMICs, more information is available on the reliability and validity of different tools; yet standardization remains challenging given the wide variety of circumstances and contexts in which they are used. Adoption of a standard tool creates the risk of misalignment to the research question, instructional practices being promoted, local capacities, or available resources, among other things, therefore jeopardizing reliability and validity (Hamre n.d.) Alternatively, a standard question bank may be a more practical way to benefit from global experience and begin to collect comparable data on classroom practices that can be used to monitor global education goals.

In sum, this review has shown that, of the admittedly limited research that attempts to relate observational measures of teaching practice to student learning outcomes, most studies have demonstrated a positive and significant association between them. However, these studies tend to rely on final ratings obtained from observational instruments; fewer studies were found that attempt to unpack the complex relationship between specific classroom practices and evidence of student learning. Such research should continue. At the same time, the technical, practical, and conceptual issues that constrain classroom observation as a means to obtain information on teaching practice,

while they are not insurmountable, are sobering and should inform measurement systems that include observational measures.

Presently, most observation criteria focus only on the teacher's skills or behaviours, not student response or impact. If we want accurate, reliable assessments of whether teachers are helping students learn, observations are needed that focus on students. More tools or additional tools are needed to examine quality, scope and relevance of the content of lessons. Some behaviours (language use, time on task) can be underreported using the snapshot method, unless intervals are very short. However, this can still provide a general picture of whether classrooms are aligned with policy and expectations.

Perhaps the most important thing that classroom observations can reliably measure that is useful for system level monitoring is time on task. It doesn't matter so much what kind of instrument is being used or the content—any classroom observation tool can and should be designed to provide a measure of whether instruction is happening in line with expectations (Are teachers present and are they using the full instructional time to introduce and review instructional concepts?) Maybe this is the only standardization we can expect, with the remainder of the questions aligned to specific project needs and capacities.

Given the available evidence and experience to date, we feel confident that the *process* of lesson and classroom observations in the context of education quality improvement programs is of great value to education stakeholders, including instructional coaches and teachers. However, concerns with reliability and validity (especially for instruments that were designed for purposes other than research) limit the use of observational data from being aggregated and reported for global, quantitative monitoring purposes. Looking back at some of the documented experiences of RTI, for example in the Philippines, Senegal and Mali, the data that showed statistically significant, quantitative correlation with EGRA scores were not the data from the timed lesson observations, but relatively more basic classroom observation checklist components.

Complexity of lesson observation, sample size considerations, time, costs, etc. all suggest that observational methods can be a powerful research tool to generate new hypotheses, and as a select impact evaluation or program monitoring tool, but there would be too many concerns with comparability and feasibility to recommend this as a globally standardized measurement tool. On the other hand, many critical factors that have been shown to be positively associated with learning outcomes can be gathered through classroom visits, and simple checklists of the classroom environment. As in other domains, system stakeholders and researchers must therefore endeavour to balance the need for standardization with the desire to understand local contextual factors that drive quality and improvement.

4 Appendix 1: Malawi EARLY grade reading assessment (EGRA) Classroom observation tool

Table 1-1: Index creation for Malawi Classroom Observation Tool

Item	Linear Weight Attributed
Teaching Preparation Index	
1. Teacher has schemes of work	0.3522
2. Teacher has a scripted lesson plan	0.2619
3. Teacher has lesson notes	0.6044
4. Teacher has assessment records	0.5447
5. Teacher has teaching, learning and assessment	0.3812
Teaching Phonics Instruction Index	
1. Teacher is able to model correct letter sounds	0.5131
2. Teacher writes letters on the chalkboard or uses print *only after* phonological awareness	0.5116
3. Teacher reviews previously learned sounds, syllables and words adding the new sounds to create words.	0.4589
4. Teacher is able to manipulate sounds by blending (putting together), segmenting (taking apart) using onsets and rimes (adding /removing initial letter to the stem)	0.5143
Teaching Reading Instruction Index	
8. Teacher teaches new words using a relevant strategy/strategies (e.g. actions, pictures, and explanation), to ensure that learners show understanding	0.4349
9. Teacher asks learners to predict story from the title and picture	0.4338
12. Teacher verifies predictions.	0.4623
13. Teacher asks comprehension questions	0.4599

14. Teacher helps learners find answers.	0.4443
General Instruction Index	
15. Teacher uses the lesson cycle, i.e. introduction, an advance organizer and the I do/We do/You do procedure, for each activity and conclusion	0.3841
16. Teacher uses appropriate pace to cater for learners with different learning and special needs	0.4112
17. Teacher varies class organization (e.g. group work, pair work, and individuals sharing work) to maximize learning	0.4251
18. Teacher supervises and supports learners through immediate and appropriate feedback	0.439
19. Teacher uses teaching, learning and assessment resources effectively	0.4009
20. Teacher uses appropriate and gender-sensitive language	0.3805
21. Teacher assigns appropriate class exercise and/ or homework	0.0663

5 Appendix 2: Sample classroom observation toolS used in Tanzania

Tanzania EdData Instruments

Table 2-1: Math classroom observation

Mathematics Classroom Observation		3	6	9	12	15	18	21	24	27	30
***Maudhui ya somo*/ Lesson Content (only one X)**											
***Namba Nzima*/Whole numbers**	*Utambuzi, kusoma na kuandika*/Identification, reading and writing										
	Kuhesabu- mamoja/Counting – ones										
	Kuhesabu- katika makundi/Counting – groups										
	Kulinganisha/Comparing										
	Kukokotoa- kujumlisha/Calculation – addition										
	Kukokotoa- kutoa/Calculation– subtraction										
	Kukokotoa- kuzidisha/Calculation – multiplication										

	Mafumbo/Word problems
***Sehemu*/Fractions**	*Kuelezea sehemu ya kitu kizima*/Describing parts of whole
	Kulinganisha/Comparing
***Jiometri*/Geometry**	*Kutaja majina ya maumbo bapa*/Naming shapes
	Kuchagua na kupanga/Classifying and sorting
	Kuchora maumbo bapa/Drawing plain figures
***Fedha*/Money**	*Utambuzi wa sarafu na noti za Tanzania*/Identifying notes and coins
	Kukokotoa (kujumlisha na kutoa) fedha/Calculating with money (addition and subtraction)
***Vitendo vya ufundishaji*/Teacher Action (only one X)**	
***Kuelezea*/Talking/explaining**	*Darasa zima kurudia maelezo ya mwalimu*/Whole class repetition/recitation
	Kuandika ubaoni/Writing on the board

	Onesho mbinu/Demonstrating
	Vielelezo na maelezo, marudio/Modeling and recitation, revision
	Kutoa kazi/Setting a task
***Maswali na majibu*/ Asking/answering questions**	*Darasa zima*/Whole class
	Vikundi vidogo vidogo/Small group
	Mmoja mmoja/Individual
***Kusaidia wanafunzi*/ Assisting pupils**	*Vikundi vidogo vidogo*/Small group
	Mmoja mmoja/Individual
Monitoring pupils and assessments	*Darasa zima*/Whole class
	Vikundi vidogo vidogo/Small group
	Mmoja mmoja/Individual
***Mwisho wa uchunguzi jibu maswali haya kwa kuzingatia ulichoona darasani*/ At the end of the observation period complete the following questions based on your general impression of the lesson.**	
Je ni vipi mwalimu ameweza kufuatilia uelewa wa wanafunzi? How well did the teacher monitor the pupils' understanding?	
Mwalimu hakuuliza swali lolote kwa wanafunzi/Teacher does not ask the pupils any questions.	

	Mwalimu aliuliza maswali ya kukumbuka na sio maswali ya kupima uelewa/Teacher asks pupils recall or repetition questions, but not questions that check for the pupils understanding (e.g. recall or repetition questions only).
	Mwalimu aliuliza maswali ya kupima uelewa, lakini hakutoa msaada zaidi/Teacher asks pupils questions to check for pupil understanding, but does not provide further assistance.
	Mwalimu aliuliza maswali ya kupima uelewa na alitoa msaada/ maelezo zaidi / Teacher asks pupils questions to check for pupil understanding and provides assistance/further explanation.
	Je ni kwa kiasi gani mwalimu aliwasaidia wanafunzi kuelewa? How well did the teacher support the pupils' understanding?
	Mwanafunzi alipotoa jibu ambalo si sahihi, mwalimu alimkaripia au kumuadhibu / When a pupil responds incorrectly, the teacher scolds or punishes the pupil.
	Mwanafunzi alipotoa jibu ambalo si sahihi, mwalimu alimtaka kujaribu tena au alimwendea mwanafunzi mwingine / When a pupil responds incorrectly, the teacher tells the pupil to try again or she moves on to another pupil.
	Mwanafunzi alipotoa jibu ambalo si sahihi, mwalimu alifafanua zaidi, alitoa vidokezo au alinyumbulisha swali katika lugha nyepesi zaidi / When a pupil responds incorrectly, the teacher asks a clarifying question, cues the pupil, or breaks down the task as appropriate.
	Hakuna jibu sahihi lililotolewa au halihusiki / No correct response given or not applicable
	Ushiriki wa wanafunzi Pupil participation
	Wanafunzi wanashiriki pale wanapotakiwa kufanya hivyo lakini si kwa kujitolea / Pupils participate when called on to do so but do not volunteer.

	Wanafunzi wanashiriki pale wanapotakiwa kufanya hivyo na wengine kwa kujitolea / Pupils participate when called on to do so and some pupils volunteer.	
	Wanafunzi nashiriki kwa bidii (pamoja na kuonesha utayari wa kuuliza na kujibu maswali, kubuni) / Pupils participate actively (including showing a willingness to ask and answer questions, make guesses.)	
	Majadiliano ya wanafunzi Pupil discussion	
	Wanafunzi hawashiriki katika majadiliano / Pupils do not engage in discussions.	
	Ushiriki wa wanafunzi umejikita katika kujibu maswali wanapoulizwa / Pupil engagement in discussions is limited to responding to questions when called on.	
	Ushiriki wa wanafunzi umejikita kwa baadhi ya wanafunzi kuanzisha mada, kuuliza na kujibu maswali wanapoulizwa / Pupils' engagement in discussion is limited to some pupils initiating topics, posing and responding to questions.	
	Wanafunzi kueleza maoni yao na kutetea hoja zao. Wanafunzi kutumia mjadala unaofaa katika kukubaliana au kutokukubaliana / Pupils state their opinions and defend them. Pupils use appropriate interaction patterns to agree or disagree.	
	Je ni kwa kiasi gani wanafunzi wameweza kujibu maswali kwa usahihi? Pamoja na: kusoma kwa ufasaha wanapotakiwa kufanya hivyo. What proportion of pupils are able to respond correctly to questions? Including: Reading with fluency when asked to read.	*Hakuna maswali yaliyoulizwa*/No questions were asked.
		Hakuna/None (0%)
		Chini ya nusu (<50%)/Less than half (<50%)
		Zaidi ya nusu (>50%)/More than half (>50%)
		Wote (100%)/All (100%)

Table 2-2: Reading classroom observation

Kiswahili Classroom Observation		3	6	9	12	15	18	21	24	27	30
***Maudhui ya somo*/Lesson Content (only one X)**											
***Kusoma*/Reading**	*Kujadili misamiati*/Discussing vocabulary										
	Kuangalia Herufi/Focus on Letters										
	Kuangalia maneno na sentensi ngazi ya: ufasaha, alama za uandishi/word and sentence level: fluency, punctuation										
	Kusoma kwa sauti (kifungu cha habari)/Reading aloud (texts)										
	Kusoma kimya (kifungu cha habari)/Reading silently (texts)										
	Kupima ufahamu wa kifungu cha habari/Checking comprehension of the text										
***Kuandika*/Writing**	*Uumbaji wa Herufi,maneno,*										

	na sentensi/Mechanical production of letters, words, and sentences										
	Kuandika utungaji/Creative writing										
	Kunakili/Copying										
	Kuchora/Drawing										
	Imla/Dictation										
***Kusikiliza*/Listening**	*Kusikiliza hadithi na maelezo*/ Listening to stories and descriptions										
	Kusikiliza sauti na maneno/Listening to sounds and words										
	Kupima ufahamu wa kifungu cha habari/Checking comprehension of the text										
	Imla/Dictation										
***Kuzungumza*/Speaking**	*Kusimulia Hadithi* / Telling/re-telling stories										
	Igizo dhima/*igizo* / Role play/drama										
	Kuimba/Singing										

	Michezo/Games										
***Sarufi* / Grammar**	*Kuelezea kanuni za lugha*/Presenting the rules of the language										
	Kufanya mazoezi ya kanuni za lugha (mazoezi ya sarufi)/ Practicing the rules (grammar exercises)										
***Vitendo vya ufundishaji*/Teacher Action (only one X)**											
***Kuelezea*/ Talking/explaining**	*Darasa zima kurudia maelezo ya mwalimu* / Whole class repetition/recitation										
	Kuandika ubaoni/Writing on the board										
	Onesho mbinu / Demonstrating										
	Vielelezo na maelezo, marudio/ Modeling and recitation, revision										
	Kutoa kazi/Setting a task										
***Maswali na majibu*/ Asking/answering questions**	*Darasa zima*/Whole class										
	Vikundi vidogo vidogo/Small group										

	Mmoja mmoja/Individual										
***Kusaidia wanafunzi*/ Assisting pupils**	*Vikundi vidogo vidogo*/Small group										
	Mmoja mmoja/Individual										
Monitoring pupils and assessments	*Darasa zima*/Whole class										
	Vikundi vidogo vidogo/Small group										
	Mmoja / Individual										

Je ni vipi mwalimu ameweza kufuatilia uelewa wa wanafunzi? How well did the teacher monitor the pupils' understanding?		
	Mwalimu hakuuliza swali lolote kwa wanafunzi/Teacher **does not ask the pupils any questions**.	
	Mwalimu aliuliza maswali ya kukumbuka na sio maswali ya kupima uelewa/Teacher asks pupils recall or repetition questions, but **not questions that check for the pupils understanding** (e.g. recall or repetition questions only).	
	Mwalimu aliuliza maswali ya kupima uelewa, lakini hakutoa msaada zaidi/Teacher asks pupils questions **to check for pupil understanding, but does not provide further assistance**.	
	Mwalimu aliuliza maswali ya kupima uelewa na alitoa msaada/ maelezo zaidi/Teacher asks pupils questions **to check for pupil understanding and provides assistance/further explanation**.	
Je ni kwa kiasi gani mwalimu aliwasaidia wanafunzi kuelewa? How well did the teacher support the pupils' understanding?		
	Mwanafunzi alipotoa jibu ambalo si sahihi, mwalimu alimkaripia au kumuadhibu/When a pupil responds incorrectly, **the teacher scolds or punishes the pupil**.	

	Mwanafunzi alipotoa jibu ambalo si sahihi, mwalimu alimtaka kujaribu tena au alimwendea mwanafunzi mwingine/ When a pupil responds incorrectly, **the teacher tells the pupil to try again or s/he moves on to another pupil**.	
	*Mwanafunzi alipotoa jibu ambalo si sahihi, mwalimu alifafanua zaidi, alitoa vidokezo au alinyumbulisha swali katika lugha nyepesi zaidi/*When a pupil responds incorrectly, **the teacher asks a clarifying question, cues the pupil, or breaks down the task as appropriate**.	
	*Hakuna jibu sahihi lililotolewa au halihusiki/***No correct response given or not applicable**	
Ushiriki wa wanafunzi Pupil participation		
	*Wanafunzi wanashiriki pale wanapotakiwa kufanya hivyo lakini si kwa kujitolea/*Pupils participate when called on to do so, **but do not volunteer**.	
	*Wanafunzi wanashiriki pale wanapotakiwa kufanya hivyo na wengine kwa kujitolea/*Pupils participate when called on to do so and **some pupils volunteer**.	
	*Wanafunzi nashiriki kwa bidii (pamoja na kuonesha utayari wa kuuliza na kujibu maswali, kubuni)/***Pupils participate actively** (including showing a willingness to ask and answer questions, make guesses).	
Majadiliano ya wanafunzi Pupil discussion		
	*Wanafunzi hawashiriki katika majadiliano/*Pupils **do not engage in discussions**.	
	*Ushiriki wa wanafunzi umejikita katika kujibu maswali wanapoulizwa/*Pupil **engagement in discussions is limited to responding to questions when called on**.	
	*Ushiriki wa wanafunzi umejikita kwa baadhi ya wanafunzi kuanzisha mada, kuuliza na kujibu maswali wanapoulizwa/*Pupils' **engagement in discussion is limited to some pupils initiating**	

	topics, posing and responding to questions.		
	Wanafunzi kueleza maoni yao na kutetea hoja zao. Wanafunzi kutumia mjadala unaofaa katika kukubaliana au kutokukubaliana/Pupils **state their opinions and defend them. Pupils use appropriate interaction patterns to agree or disagree**.		
	Je ni kwa kiasi gani wanafunzi wameweza kujibu maswali kwa usahihi? *Pamoja na: kusoma kwa ufasaha wanapotakiwa kufanya hivyo.* **What proportion of pupils are able to respond correctly to questions?** Including: Reading with fluency when asked to read.	*Hakuna maswali yaliyoulizwa*/No questions were asked	
		Hakuna/None (0%)	
		Chini ya nusu (<50%)/Less than half (<50%)	
		Zaidi ya nusu (>50%)/More than half (>50%)	
		Wote (100%)/All (100%)	

6 Appendix 3: Abbreviations and Acronyms

CLASS	Classroom Assessment Scoring System
EDC	Educational Development Center
EDDATA	Education Data for Decision Making
EGMA	Early Grade Mathematics Assessment
EGR	early grade reading
EGRA	Early Grade Reading Assessment
FFT	Framework for Teaching
GPS	global positioning system
LMIC	low- and middle-income country
MELQO	Measuring early learning and quality outcomes
MET	Measures of Effective Teaching project
MoEST	Ministry of Education, Science and Technology
MTB-MLE	mother-tongue-based multilingual education curriculum
OLS	ordinary least squares
OPEQ	Equitable Access to Quality Basic Education
ORF	oral reading fluency
PEA	Primary Education Advisor
RARA	Nigeria Reading and Access Research Activity
SD	standard deviation
SDG	Sustainable Development Goal
SSME	School Snapshot of Management Effectiveness
TAC	Teacher Advisory Centre
TIPPS	Teacher Instructional Practices and Processes System
UNESCO	United Nations Educational, Scientific and Cultural Organization
USAID	US Agency for International Development

7 Bibliography

Aaronson, D, Barrow, L, & Sanders, W 2003, 'Teachers and student achievement in Chicago public high schools'. Working paper 2002-28. Federal Reserve Bank of Chicago, Chicago, IL.

Akiva, T, & Li, J 2016, 'Participation', figure, in *The Simple Interactions Tool*, Simple Interactions, viewed 14 April 2016 http://www.simpleinteractions.org/the-si-tool.html

Al Otaiba, S, Clancy-Menchetti, J, & Schatschneider, C 2006 'Examining the effects of professional development to improve early reading instruction: How strong is the causal chain?', in Scruggs, T., and Mastropieri, M. (eds.) *Applications of Research Methodology,* vol. 19, pp. 201–236, Emerald Group Publishing Limited, Bingley, UK.

Aslam, M, & Kingdon, G. 2012, 'How teachers' pedagogic practice influences learner achievements: A study in the Punjab, Pakistan'. In Moon, B. (ed.). *Teacher education and the challenge of development: A global analysis,* Routledge, New York.

Ballou, D, Sanders, W, & Wright, P 2004, 'Controlling for student background in value-added assessment of teachers', *Journal of Educational and Behavioral Statistics,* vol. 29, no. 1, pp. 37–65.

Bommer, WH, Johnson, JL, Rich, GA, Podsakoff, PM, & MacKenzie, SB 1995. 'On the interchangeability of objective and subjective measures of employee performance: A meta-analysis', *Personnel Psychology,* vol. 48, no. 3, pp. 587–605.

Boyd, D, Grossman, P, Lankford, H, Loeb, S, and Wyckoff, J 2006. 'How changes in entry requirements alter the teacher workforce and affect student achievement', *Education Finance and Policy*, vol. 1, pp. 176–216.

Brombacher, A, Nordstrum, L, Davidson, M, Batchelder, K, Cummiskey, C, & King, S 2014, *National baseline assessment for the 3Rs (reading, writing, and arithmetic) using EGRA, EGMA, and SSME in Tanzania*, USAID Office of Sustainable Development, Tanzania.

Brown, JL, Jones, SM, LaRusso, MD, & Aber, JL 2010, 'Improving classroom quality: Teacher influences and experimental impacts of the 4Rs program', *Journal of Educational Psychology*, vol. 102, no. 1, pp. 153–167.

Clotfelter, CT, Ladd, HF, & Vigdor, JL 2006. 'Teacher-student matching and the assessment of teacher effectiveness', *Journal of Human Resources,* vol. 41, no. 4, pp. 778–820.

Cohen, J 1992. 'A power primer', *Psychological Bulletin,* vol. 112, no. 1, pp. 155–159.

Cuban, L 2013. *Inside the black box of classroom practice: Change without reform in American education*. Cambridge, MA: Harvard Education Press.

Danielson, C 2011, *The Framework for Teaching evaluation instrument,* The Danielson Group, Princeton, NJ.

Education Development Center, Inc. 2013. *USAID/Philippines BASA Pilipinas Program Annual Progress Report, January 1–December 31, 2013* http://pdf.usaid.gov/pdf_docs/PA00JXG4.pdf

Gallagher, HA 2004, 'Vaughan Elementary' s innovative teacher evaluation system: Are teacher evaluation scores related to growth in student achievement', *Peabody Journal of Education,* vol. 79, no. 4, pp. 79–107.

Garrett, R, & Steinberg, MP 2015. 'Examining teacher effectiveness using classroom observation scores: Evidence from the randomization of teachers to students', *Educational Evaluation and Policy Analysis,* vol. 37, no. 2, pp. 224–242.

Geddes, M 2015. *What constitutes a Global Rating Scale* http://www.getyardstick.com/what-constitutes-a-global-rating-scale/

Goldhaber, DD, & Brewer, DJ 1997. 'Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity', *The Journal of Human Resources,* vol. 32, no. 3, pp. 505–523.

Hamre, BK n.d. *Using classroom observations to gauge teacher effectiveness. The Classroom Assessment Scoring System (CLASS)*, PowerPoint presentation http://www.gtlcenter.org/sites/default/files/docs/Hamre.pdf

Harris, DN 2012, *How do value-added indicators compare to other measures of teacher effectiveness?*, The Carnegie Foundation for the Advancement of Teaching, Carnegie Knowledge Network, Stanford, CA.

Harris DN 2013, *How might we use multiple measures for teacher accountability?*, The Carnegie Foundation for the Advancement of Teaching, Carnegie Knowledge Network, Stanford, CA.

Harris, DN, & Herrington, CD 2015. 'Editors' introduction: the use of teacher value-added measures in schools: New evidence, unanswered questions, and future prospects', *Educational Researcher,* vol. 44, no. 2, pp. 71–76.

Harris, DN, & Sass, TR 2007. *Teacher training, teacher quality, and student achievement*. Working paper 3. National Center for the Analysis of Longitudinal Data in Education Research, Washington, DC.

Harris, DN, & Sass, TR 2009. *What makes for a good teacher and who can tell?* Working paper 30. National Center for the Analysis of Longitudinal Data in Education Research, Washington, DC.

Heneman, RL 1986. 'The relationship between supervisory ratings and results-oriented measures performance: A meta-analysis', *Personnel Psychology,* vol. 39, no. 4, pp. 811–826.

Ho, AD, & Kane, TJ 2013, *The reliability of classroom observations by school personnel*. The Bill & Melinda Gates Foundation, Seattle, WA.

Hoxby, CM 2002, 'Would school choice change the teaching profession?' *Journal of Human Resources,* vol. 37, no. 4, pp. 846–891.

Jacob, B, & Lefgren, L 2008, 'Can principals identify effective teachers? Evidence on subjective performance evaluation in education', *Journal of Labor Economics,* vol. 26, no. 1.

Jepsen, C 2005, 'Teacher characteristics and student achievement: Evidence from teacher surveys', *Journal of Urban Economics,* vol. 57, no. 2, pp. 302–319.

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

Kantor, H, & Lowe, R 2004, 'Reflections on history and quality education', *Educational Researcher,* vol. 33, no. 5, pp. 6–10.

Kimball, SM, White, B, Milanowski, AT, & Borman, G. 2004, 'Examining the relationship between teacher evaluation and student assessment results in Washoe County', *Peabody Journal of Education,* vol. 79, no. 4, pp. 54–78.

Kingdon, G 1996, *Student achievement and teacher pay: A case-study of India,* STICERD Discussion Paper No. 74. London School of Economics. London, UK.

Measures of Effective Teaching Project 2012, *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. The Bill & Melinda Gates Foundation, Seattle, WA.

Milanowski, AT 2004, 'The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati', *Peabody Journal of Education,* vol. 79, no. 4, pp. 33–53.

New York University 2016, *OPEQ: Opportunities for equitable access to quality basic education in the Democratic Republic of Congo*, viewed 26 March 2016. http://steinhardt.nyu.edu/ihdsc/opeq/tipps

Nordstrum, LE 2015, *Effective teaching and education policy in sub-Saharan Africa: A conceptual study of effective teaching and review of educational policies in 11 Sub-Saharan African countries,* USAID, Washington, DC and RTI International, Research Triangle Park.

Pianta, RC, & Hamre, BK 2009, 'Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity', *Educational Researcher,* vol. 38, no. 2, pp. 109–119.

Podgursky, M, & Springer, M 2007, 'Credentials versus performance: Review of the teacher performance pay research', *Peabody Journal of Education,* vol. 82, no. 4, pp. 551–573.

Pouezevara, SR, DeStefano, J, Cummiskey, CP, & Pressley, J 2015, *Philippines EGRA four language study: 2015 follow-on*. RTI International, Research Triangle Park, NC, viewed 21 June 2016 https://www.eddataglobal.org/countries/index.cfm?fuseaction=pubDetail&ID=831

Raudenbush, SW, & Jean, M 2012, *How should educators interpret value-added scores?* The Carnegie Foundation for the Advancement of Teaching, Stanford, CA.

Rivkin, SG, Hanushek, EA, & Kain, JF 2005, 'Teachers, schools and academic achievement', *Econometrica*, vol. 73, no. 2, pp. 417–458.

Rockoff, JE 2004, 'The impact of individual teachers on student achievement: Evidence from panel data', *American Economic Review,* vol. 94, no. 2, pp. 247–252.

Rockoff, JE, & Speroni, C 2010, 'Subjective and objective evaluations of teacher effectiveness', *American Economic Review, Papers and Proceedings*, vol. 100, no. 2, pp. 261–266.

RTI International 2015, *Early Grade Reading Assessment (EGRA) national baseline assessment in Mali*, USAID, Washington, DC.

RTI International 2016, *Nigeria Reading and Access Research Activity (RARA): Results of an approach to improve early grade reading in Hausa in Bauchi and Sokoto States*. USAID, Abuja, Nigeria, viewed 21 May 2016 https://www.eddataglobal.org/countries/index.cfm?fuseaction=pubDetail&ID=846

RTI International 2016, Early Grade Reading Assessment (EGRA) Toolkit: Second Edition, USAID, Washington, DC, viewed 5 July 2016. https://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&id=929

Stallings, J 1977, *Learning to look: A handbook on classroom observation and teaching models*. Wadsworth Publishing Company, Belmont, CA.

Stuhlman, MW, Hamre, BK, Downer, JT, & Pianta, RC n.d., 'How to select the right classroom observation tool', In *A practitioner's guide to conducting classroom observations: What the research tells us about choosing and using observational systems to assess and improve teacher effectiveness.* The Center for Advanced Study of Teaching and Learning, University of Virginia, Charlottesville, VA, viewed on 10 May 2016 http://curry.virginia.edu/uploads/resourceLibrary/CASTL_practioner_Part3_single.pdf

The New Teacher Project 2011, *Rating a teacher observation tool: Five ways to ensure classroom observations are focused and rigorous.* PowerPoint presentation, viewed on 1 May 2016 http://tntp.org/assets/documents/TNTP_RatingATeacherObservationTool_Feb2011.pdf

Texas A&M University 2001, *Stallings snapshot observation manual. January 2007. Modified for use in The World Bank projects,* Center for Collaborative Learning Communities, College Station, TX.

UNESCO 2004, *Education for all global monitoring report 2005: The quality imperative*. UNESCO Publishing, Paris.

UNESCO Technical Advisory Group 2015, *Technical Advisory Group proposal: Thematic indicators to monitor the post-2015 education agenda*. World Education Forum, UNESCO Publishing, Paris.

Varly, P 2010, *The monitoring of learning outcomes in sub-Saharan Africa: Senegal. Languages of instruction and teachers' methods in Senegal Grade 3 classrooms*. William and Flora Hewlett Foundation Grant, Menlo Park, CA, viewed 15 May 2016 https://www.eddataglobal.org/countries/index.cfm?fuseaction=pubDetail&ID=349

Varma, A, & Stroh, LK 2001, 'The impact of same-sex LMX dyads on performance evaluations', *Human Resource Management,* vol. 40, no. 4, pp. 309–320.

Vegas, E, & Umansky, I 2005, 'Improving teaching and learning through effective incentives', In Vegas, E. (ed.) *Incentives to improve teaching*. The World Bank, Washington, DC.

Waxman, HC, Dubinski Weber, N, Franco-Fuenmayer, S, & Rollins, KB 2015, 'Research-based approaches for identifying and assessing effective teaching practices: Challenges, new directions and policy implications', In Li, Y., and Hammer, J. (eds.) *Teaching at work*. Sense Publishers, Rotterdam, NY.

Westbrook, J, Durrani, N, Brown, R, Orr, D, Pryor, J, Boddy, J, & Salvi, F 2013, *Pedagogy, curriculum, teaching practices and teacher education in developing countries. Final report. Education rigorous literature review*. EPPI-Centre, Social Science Research Unit, Institute of Education, University of London, London, UK.

Whitehearst, GJ, Chingos, MM, & Lindquist, KM 2014, *Evaluating teachers with classroom observations: Lessons learned in four districts.* Brown Center on Education Policy at Brookings Foundation, Washington, DC.