Technical note on methodology and analytics for the Fursa kwa Watoto (FkW) Learning Agenda

A. Introduction

The Fursa kwa Watoto (FkW) Collaborative developed and tested a pre-primary package of interventions in an effort to establish a low-cost model of quality education in Tanzania. The project approach, components, and all results are available online at http://www.fkwlearningagenda.com. In the **Learning Agenda**, we used research methods to monitor, evaluate and learn about the implementation, outcomes and impacts of FkW in a changing context. The evaluating organizations in the FkW collaborative (Mathematica and Centre for Social Responsibility (CSR) Group Africa) conducted a range of activities including a randomized control trial (RCT) of impacts on student learning, repeated observations of teachers' instructional practices, school finances, and student enrollment and attendance. More specifically, Learning Agenda activities include:

1) An assessment of learning outcomes among 1,229 pre-primary students conducted at three time points (two in pre-primary, one at the end of standard 1) in intervention (n=65) and control schools (n=66)
2) Six rounds of classroom observations in 100 schools (80 expansion and 20 pilot schools)
3) Telephone surveys with head teachers (n=130) to collect monthly enrollment and attendance data, and capitation grant and parent and other contribution data
4) Qualitative in depth interviews (n=170 in total over two times points) with teachers (n=40), paraprofessionals (n=18) head teachers (n=40), ward (n=8), quality assurance (n=12), district education (n=4) and executive officers (n=4),
5) Focus group discussions (FGDs) (n=40 in total over two time points) with School Management Committees (SMCs) (n=16), community members (n=12), and parents (n=16), and
6) A costing study using program data.

This technical note describes the study design, sampling, instrumentation, training procedures, and analytic approach to each of the study components.

B. Methods and approaches

We implemented the following methods to answer the study's research and evaluation questions. Below we list each component of the study and describe our approach to each subcomponent:

1. Sampling for all study components

In 2015, we implemented a school-mapping exercise in four districts, two in Mwanza and two in the Kilimanjaro regions of Tanzania. CSR visited and mapped schools. All schools that had not participated in the FkW pilot and were accessible to the FkW implementers (within a two hour drive) were eligible. The mapping process yielded basic statistics on the school, school leadership, pre-primary teachers, and students. School information included the number of pre-primary teachers, resources allocated to pre-primary education, and the distance from the school to a central landmark, such as the district center. Teacher information included pre-primary teachers' qualifications, years of teaching, age, and gender. Student information included Standard VII leaving exam scores for the latest available year, the number and ages of students, and students' primary language. Following the mapping, we implemented the following procedures to randomize schools to an intervention or control group for the expansion stage of FkW and to ensure balance between groups assigned to either intervention or control status:

1. In early 2016, using the 2015 mapping data, we excluded schools without pre-primary classrooms; schools where teachers weren't willing to participate; and schools located in Ilamela, which was deemed too far away for implementers to reach because it is located over 100 kilometers from Mwanza and accessible only along poor quality roads.
2. We created 11 strata by region and district (Misungwi, Nyamagana districts in Mwanza and Moshi rural, and Moshi urban districts in Kilimanjaro), and by performance based on standard 7 exam scores (Table 1). Performance was rated as low, medium, or high. In the urban region of Moshi, schools with low and medium performance were grouped together due to the small number of schools. The table below shows how strata were allocated among regions, districts, and student performance.

Table 1. Distribution of strata across the regions and districts based on student performance

Stratum		Student Performance		
	District	Low	Medium	High
Eligible schools in Mwanza	Misungwi	1	2	3
	Nyamagana	4	5	6
Eligible schools in Kilimanjaro	Moshi rural	7	8	9
	Moshi urban	11		11

3. Next, we selected schools across regions and districts—proportional to the size of the strata—to reach a sample of 240 schools (*the original sample size)*. We then randomized schools from each stratum into intervention (n = 120) and control groups (n = 120) using a random number. Next, we assessed balance on several variables, such as number of pre-primary teachers and pre-primary enrollment.

   Note that, at this stage, the 120 intervention schools became "expansion schools" for the next stage of the FkW initiative. The implementing partners, Children in Crossfire (CiC), Aga Khan University (AKU), Maarifa, and TAHEA, began implementing the FkW training and package of services including Component 1, Model 1; Component 2 at the level of the district, ward, school management committee, and schools; and Component 3 at the national level.

4. Next, in 2017, we used this larger group of 240 intervention and control schools to select the sample for most of the Learning Agenda activities. Note that the Steering Committee agreed to reduce the sample size for the student assessment so that resources could be reallocated to the classroom observations and the qualitative portions of the study.

   For the student assessment, we reduced the sample size from 240 schools to 130 schools (65 intervention schools and 65 comparison schools), with an even distribution across Mwanza and Kilimanjaro. Schools were selected proportionally by stratum based on the original assignment from the larger sample of 240 schools. Table 2 shows the number of schools selected from each stratum for the intervention and comparison groups, respectively. We used this sample of 130 schools for the student assessment and the study of student enrollment and attendance. (Table 2).

5. Further, for the student assessment, in the 130 schools we implemented a process to randomly select 12 students per school. Our field team worked with teachers to group students by age. We listed the children's ages and randomly selected 12 students—ages 5 or 6—to participate in the assessment. If the student refused to participate, we selected a replacement. Several times, in order to reach our target of 12 students, we had to include a seven-year-old. We conducted student assessments in May 2017, November 2017, and November 2018.

6. For the classroom observation study, we further selected a reduced sample of 80 schools (from this larger sample of 130 expansion schools) in which to conduct observation study.[1]Again, we ensure baseline equivalency. We conducted classroom observations in May 2017, November 2017, March 2018, and November 2018.

7. For the qualitative study, we based the sample size on the number of interviews or focus groups we believed would be necessary to reach saturation across the study groups. Then we randomly selected the location and the schools from which we would invite respondents to participate (Table 3). Round one of qualitative data collection was completed in October 2017 and round two in October 2018. Note that if round one participants could not participate in round two, we selected their replacement in the same location or school.

Table 2. Distribution of schools for the student assessment (n = 130)

		Student performance			
Intervention schools	District	Low	Medium	High	Total
Mwanza	Misungwi	8	9	4	
	Nyamagana	2	4	5	32
Kilimanjaro	Moshi rural	11	7	9	
	Moshi urban	2		4	33
				Total	**65**
Comparison schools					
Mwanza	Misungwi	8	9	4	
	Nyamagana	2	4	5	32
Kilimanjaro	Moshi rural	11	7	9	
	Moshi urban	2		4	33
				Total	**65**

Note that we administering the student assessment tool to children and teachers in a *late* baseline in May 2017 and at the end of the school year in November or December 2017.

Table 3. Qualitative sample

		Distribution					
		Moshi			**Mwanza**		
		Intervention	**Control**	**Pilot**	**Intervention**	**Control**	**Pilot**
Interviews	Teachers	8	8	4	8	8	4
	Paraprofessionals	3	3	3	3	3	3

[1] In addition to the expansion schools, we added 20 pilot schools for the classroom observation study as well. See section on classroom observation study.

	Head teachers	8	8	4	8	8	4
	DAO	2			2		
	DEO	2			2		
	WEO	2	2		2	2	
	VEO						
	QAO	3	3		3	3	
	Sub Total	28	24	11	28	24	11
FGDs	Parents	3	3	2	3	3	2
	SMCs	3	3	2	3	3	2
	Community	2	2	2	2	2	2
	Sub Total	8	8	6	8	8	6
	TOTAL	**36**	**32**	**17**	**36**	**32**	**17**

C. Study activities

Next we describe each of the study components including the purpose of the evaluation activity, the tools used, the training process, and analytical procedures. For each study, additional information is available upon request.

1. Student assessments

The purpose of the student assessment was to collect preliminary evidence on student learning and development and explore whether FkW led to improved outcomes among pre-primary students. We designed a randomized control trial (RCT) to measure differences in student outcomes. The collaborative recognized that the context of pre-primary became more challenging with increased enrollment and a persistent teaching shortage. Still, given the value of acquiring preliminary evidence on student progress throughout the school year, we decided to use the RCT design and conducted a student assessment at three time periods to compare outcomes based on the intervention status of the school and school characteristics. This study provides data on the impacts of FkW, and insight into student's pre-academic skills, social development, and executive function. Note that we were unable to conduct a subgroup analysis with statistical significance given the sample size, yet this study does provide high quality exploratory data on students' foundational skills over time.

a. Student assessment tool

We used the National Pre-Primary Curriculum and Syllabus (2016) and the Basic Education Syllabus for Standard 1 (2018) to guide our selection of assessment tools. First, we assessed a cohort of pre-primary students using the Measuring Early Learning Quality and Outcomes (MELQO) tool. The MELQO Consortium—which includes UNESCO, UNICEF, the World Bank, The Brookings Institute, the Global Partnership for Education, and the World Health Education—developed and validated the tool. The tool had been used in a national study across Tanzania and had gained acceptance among educations stakeholders. The MELQO takes about 35 minutes to administer and can be used with children ages 3 to 6 years. The child assessment includes a set of 25–30 core items drawn from existing regional and international tools and was designed to assess child development and learning. The tool assesses pre-literacy, pre-numeracy, socio-emotional skills, and areas that support learning across multiple domains, such as executive function, persistence, and self-regulation. We used the MELQO at baseline in May and midline in November 2017.

By November 2018, most students—now a year older—had transitioned to Standard 1. We updated our assessment to reflect students' advancing skills. We used items from the Early Grade Reading Assessment (EGRA) and the Early Grade Math Assessment (EGMA). MELQO items in which students already scored high and had limited room for growth were dropped from the assessment. Guided by the Basic Education Syllabus for Standard 1, and with input from Standard 1 teachers, we selected similar, but slightly more challenging items from the Early Grade Reading and Math Assessments to add to our tool. The assessment tool was pretested in Mwanza in October 2018 and subsequently finalized. Note that we only assessed outcomes aligned with the Basic Education Syllabus for Standard I, though we did not assess every outcome within the syllabus due to time constraints with each child.

b. Training and field procedures

At each data collection stage, the field team of assessors and supervisors trained and piloted the tools over five to eight days in Mwanza or Kilimanjaro. After two days of introduction to the tool and practice, the team practiced the tool by conducting the assessments in teams with children at seven schools. The teams of three observers jointly completed one assessment with each child. Following the assessment, observers individually entered the data into a tablet without discussing the student's performance with other observers. Once all data were entered each day, we examined the inter-rater reliability (IRR) for the pilot MELQO data between our observers. We debriefed on the assessments and methodically reviewed the IRR data item by item to understand agreement and disagreement in observer ratings. We clarified the questions, responses, and definitions related to items that had low reliability in order to improve overall reliability across enumerators. Across the piloting, we achieved an average IRR of 96 percent. Data collection was conducted in May and early June 2017 for the baseline, in November 2017 for the midline, and November 2018 for the endline.

c. Analytic plan

Scoring: Enumerators followed standard scoring procedures outlined in the MELQO training manuals published by the MELQO consortium. However, while the assessment and its administration standardized for use across a large number of countries, MELQO data has not, to our knowledge, been used as an assessment as part of a longitudinal impact evaluation. Further the tool was not scaled or adjusted to contain more difficult items from one assessment to the next. In the absence of other validated tools for this population in Tanzania, and given the age of students and their level of performance at baseline, we determined that it was not too serious a risk to the study to repeat the same instrument at baseline and midline. However, in order to use the MELQO assessment data to meaningfully compare two groups and assess students' growth over time, we adapted the scoring mechanism to include all students, even those who were subject to the "stop rule" or unable to provide a response. Therefore, students who were subject to the "stop rule," or provided no response were treated as 0, or incorrect scores. We produced summary scores by grouping like items at the domain and the skill level, and out of these items, calculated the percent of total points out of all total possible points.

Analysis: Next, we analyzed the outcomes across each task. For the baseline MELQO student assessment data, we cleaned the data and calculated descriptive statistics to examine students' pre-academic skills, including language, pre-literacy and pre-numeracy, socio-emotional skills, and areas that support learning across multiple domains, such as executive function, persistence, self-regulation, and approaches to learning.

We calculated regression-adjusted means and adjusted by the strata used to select the sample for baseline and midline samples. For the MELQO data, the means provide a snapshot of changes over time, and differences between groups. We explored differences based on intervention status, the location of the school, enrollment size, and teacher characteristics such as years of teaching and certification. We also explored differences in outcomes based on students' characteristics, including age and gender. Because we conducted the student assessment in May and not at the beginning of the school year in February, we cannot eliminate the possibility that students demonstrated different levels of performance before our assessment. However, we employed a strong sampling approach to increase the likelihood of observing baseline equivalency between schools. In theory, all students, regardless of study group, would have a similar learning trajectory so that growth above and beyond that "normal" learning can be attributed to FkW.

For the endline student assessment, we repeated the above procedures. Again, we cannot be fully confident that there was baseline equivalence between the two groups of students, so we must be cautious in attributing differences in students' abilities' to FkW. However we note that we can trace back the few impacts we found to the instructional practices of teachers and impacts are corroborated in qualitative interviews, thus we believe this study provides preliminary evidence at the student level.

We define the impact of FkW as the difference between the intervention and comparison groups in students' average outcomes at endline. We explore differences between students based on FkW using a DID approach that compares changes between baseline and endline for students in FkW schools over time. Specifically, we used the following regression framework:

$$(1)\ y_{ijt} = \propto + Post_t + \pi Interv_j \ + \rho Interv_j * Post_t \ + \varphi X_{ij} \ + \vartheta Z_j + \mu_j + \varepsilon_{ijt}$$

where y_{ijt} is the outcome of interest for student *i* in school j in time *t*; $Post_t$ is a dummy variable where "1" represents the post-intervention period; $Interv_j$ is a binary variable equal to "1" if the school was assigned to receiving FkW and zero otherwise; X_{ij} and Z_j are vectors of baseline student- and school-level characteristics, respectively, that can affect the outcome of interest but are unrelated to the project (for example, students' scores at baseline; gender; or school location); μ_j is a school-specific random error term; and ε_{ijt} is a student-specific random error term.

The parameter of interest in Equation (1) is ρ, the DID estimate, which is an estimate of the average impact of an FkW school adjusting for other factors. This is an intent-to-treat estimate because not all students took advantage of the program (for example, students might have attend classes infrequently). Therefore, it can be interpreted as the effect of attending an FkW assigned school. Because the unit of intervention is the school, we accounted for the correlation in outcomes among students in the same school, district, and region when estimating the standard error for the estimate ρ. We clustered the standard errors to account for the nesting of students in schools and include dummy variables to represent strata used in random assignment. Our main model included the intervention indicator and strata dummies.

To assess the robustness of our conclusions we conducted sensitivity analyses that included student- and school-level covariates such as student gender, grade at endline (about 11 percent of students were held in pre-primary), student performance in the baseline assessments, pupil-teacher ratios, and the region where the school is located (Mwanza and Kilimanjaro). We also ran the analyses separately by region and excluding students who were held in pre-primary school. The conclusions are generally robust across models, so in the final results we presented the most parsimonious model.

Attrition: As with all interventions, not all students take advantage of the program (for example, students might attend classes infrequently), thus results can be best interpreted as the effect of attending an FkW assigned school. Since the study is using change over time to assess program impacts, the results of the student analysis may reflect factors other than FkW if students who left the sample versus those who remained were different in the two study groups. Therefore, we conducted a simple **attrition analysis** to assess whether attrition from the sample might affect these findings. According to teachers, the most common reason for attrition was that the student's family moved to another area, followed by illness, transfers to other schools, and absenteeism. We found that students in the intervention group were *not* more likely than control students to leave the sample before the midline assessment or the endline assessment. Based on the relatively low amount of sample loss and minimal differences between groups, attrition is unlikely to have biased study results or explain any differences in student scores at midline according on What Works Clearinghouse standards (a website that reviews the quality of educational studies in the United States).[2]

2. Classroom observation

a. Description and design

The classroom observation tool and process was designed to provide detailed insights into instructional practices and learning environments across a range of dimensions over time, based on schools and teachers participating in FkW. The observations allow us to measure and assess teachers' instructional practices, behaviors, and methods that are along the causal pathway between FkW training and student learning. The longitudinal approach enables an assessment of whether teachers who participate in FkW, both certified teachers and paraprofessionals, are in fact taking up the intervention as described in the theory of change—compared to teachers who have not participated in FkW—and whether they continue to implement the practices as they receive ongoing coaching. Although the links between training, instructional practices, and student learning are critical underpinnings to the theory of change for most in-service training and professional development programs, a literature review revealed relatively few rigorous evaluations that test these links or that test this overarching theory of change, especially in developing countries and among pre-primary teachers. This study relying on repeated teacher observations can help begin to fill an important gap in this literature.

b. Sample

Continuing from the earlier description of the study sample, for the classroom observations, we selected a sample of 80 expansion schools (as mentioned) and a sample of 20 FkW pilot schools for the classroom observation activity. The expansion schools participated in FkW between 2016 and 2017 if they were randomized into the intervention group, whereas control group schools did not participate in FkW. Pilot schools participated in FkW training and mentoring activities between 2014 and 2015. We implemented the following procedures:

1. First, for observations in the expansion schools, we used the same procedures and approach that we implemented to select the schools for the MELQO assessment. The only difference is that we

[2] The What Works Clearinghouse (WWC) is an initiative of the U.S. Institute of Education Sciences to evaluate studies on the effectiveness of programs, policies and practices. WWC Standards Briefs lay out rules to assess the quality of studies and are highly regarded in the field of program evaluations.

planned to conduct the observations in 40 intervention and 40 control schools, rather than 65 schools in each study group. We randomly selected 40 intervention and 40 control schools across the strata from the larger group of intervention and control schools that had been previously selected. The sample includes 20 intervention and 20 control schools in both regions, for a total of 40 schools in Moshi and 40 schools in Mwanza.

2. Second, for observations in the pilot schools, we randomly selected 20 schools that participated in the pilot study in 2014 and 2015. There was no control group in this sample. To select the sample, we grouped schools into three strata based on an earlier assessment of teachers' instructional practices by AKU and the TWG during the FkW pilot. At that time, teachers in pilot schools were assigned to categories of "strong," "average," and "weak," based on multiple assessments of their instructional practices. We selected several schools in each of these categories in both Moshi and Mwanza to observe at two time points in 2017. The sample includes schools in each stratum, for a total of 10 schools in Moshi and 10 schools in Mwanza.

c. Tool

The FkW Steering Committee partners developed the classroom and teacher observation tool and rubric in a collaborative and iterative process beginning in 2015, with a finalized tool developed in 2016. This first iteration of the tool was developed by AKU as a way to assess teachers' instructional practices, the classroom learning environment, and other factors related to the AKU teacher training course. AKU administrators assessed the tool's face validity and approved the tool internally. At the same time, the FkW Technical Working Group (TWG), including staff at Maarifa and Tahea, developed and began implementing a second tool designed to capture concepts related to the learning environment that were not otherwise measured by the first tool, which focused more on instructional practices. Mathematica led the integration of these two tools, followed by approval from AKU and the TWG. This combined observation tool was used by AKU during classroom visits throughout 2016 and 2017. The tool assesses the quality of the learning environment and teacher performance in the following areas:

- Organization of the school day
- Lesson plan development and use
- Instructional strategies and skills
- Use of learning materials and classroom resources
- Appropriateness, quality, and quantity of learning materials
- Children's participation in learning
- Teacher interaction during play sessions
- Classroom management

For the classroom observations in the Learning Agenda, we added several items from the MELQO Classroom Observation Form to the latest version of the FkW Classroom Observation Tool. For example, we added items on specific instructional practices that teachers implemented during pre-writing, pre-reading, and pre-numeracy activities. We also added items on the school environment, such as physical space—both indoors and outdoors—water source, handwashing and toilet facilities, and feeding programs.

d. Training and field procedures

For all data collections, CSR Group Africa led the training with oversight from Mathematica. At baseline, the training entailed three days of closely reviewing and practicing with the classroom observation tool, followed by three days of piloting the tools in the morning and debriefing on the piloting process during the afternoon. Subsequent trainings for the teacher observation training included two days of full-time training, followed by four days of piloting.

We observed the full pre-primary session, from the beginning of circle time to the closing activities. We then conducted a post-pilot briefing section to discuss challenges encountered during the pilot. After each observation, we assessed the pilot data for IRR among the three observers who visited a given classroom as a measure of consistency across evaluators' judgments. We wanted to ensure that enumerators had a common understanding of the classroom observation tools, and scored teacher instruction in the same way across all study subjects. High inter-rater reliability, coupled with intensive training on how to interpret performance and instruction increased our confidence that enumerators were scoring subjects consistently and accurately. We clarified and discussed the questions, responses, and definitions for items that had low reliability in order to improve overall reliability across enumerators. The IRR scores entailed calculating the percent of agreement for each item for all coders who evaluated a single student or classroom. Items where there was significant disagreement, enumerators were asked to identify their reasons for scoring in a particular way, and enumerators and training staff worked together to develop a consensus on the appropriate code based on specific evidence from the class and training materials. Through this process, IRR increased over the course of training. Piloting continued until all enumerators exceeded the minimum standard of at least 80 percent agreement and reached 95 percent.

Baseline data collection in the expansion schools was conducted in May 2017 while data collection in the pilot schools was conducted in July 2017 because those schools closed for most of June. Observations were then conducted in November 2017, March 2018 and November 2018.

e. Analytic plan

Scoring: We focused on FkW components such as lesson planning, instructional skills, learning materials, student participation, and classroom management. The tool also captures aspects of the school environment, including school feeding and sanitation facilities. Teachers received a score from 1 (poor) to 5 (excellent) in each domain. Scores were then converted to percentages with 1=20% and 5=100%. Additionally, items from the same section of the tool were grouped, and summary scores were produced by calculating the percent of points received out of all possible points.

Analysis: For the first round of data collection, we calculated descriptive statistics and, whenever possible, compared mean composite scores to examine instructional practices and strategies, organization of the school day, the classroom environments and use of learning materials, children's participation, and classroom management. Given that we are unable to observe these teachers before training, we cannot know whether teachers demonstrated different levels of effectiveness before training, however we do know that there was baseline equivalence between the schools as we had assessed balance during the sampling process. We note that differences in instructional practices can be clearly traced back to the intervention and are strongly supported by qualitative reports.

For the second and subsequent rounds of data collection, we replicated the first analytical plan, but continue the exploration to understand how instructional practices and the classroom environment change over the course of the school year. We explored whether teachers' practices progressed, remained steady, or regressed, and examined differences based on the intervention status, ongoing activities at the ward or district level or implemented by parents, and characteristics of the schools and teachers. For the pilot school data, we compared early baselines conducted in 2014 to observations conducted at the end of 2015, along with the observations conducted in May 2017, November 2017, and March and November 2018 to assess sustainability of practices and the environment.

We calculated regression-adjusted means for all data collected, including enrollment and attendance. Results were adjusted by the strata used to select the sample. For classroom observations, because (1) the baseline was implemented soon after training, and (2) whereas teacher training programs take time to affect learning outcomes, teachers can implement new practices immediately. Therefore, we feel comfortable attributing the large, early differences to FkW rather than preexisting differences between schools. We developed tables and figures that illustrate differences in average scores by time period, location, and study group.

3. Enrollment and financial analysis

a. Description, tool, and field procedures

The purpose of the enrollment study was to track changes in students' enrollment and attendance at multiple time points in the expansion schools. The financial analysis was designed to understand how much schools receive per month from the government in the capitation grant allotment and from family and other contributions.

When we visited schools to conduct the MELQO student assessments and student observations, we discuss the enrollment and the financial study with the headmaster. We collected enrollment and attendance data by age, gender, and student's language from study schools. For the first round of data collection, our field team asked the headmaster to review the school's pre-primary student listing. Working with the teacher, the field team determined the gender, age, and first language of the students. We also collected monthly financial data. Head masters, with approval, shared monthly grant information as well as monthly contribution data.

For each round of data collection, we used a template that captures latest statistics on pre-primary students and financial grants and contributions. We optimized our use of resources by collecting data during the student assessment visits and then also telephoned head teachers and classroom teachers to collect updated information. We repeated data collection at regular intervals (March, September, and November 2017, May, September, and November 2018). We made follow up phone calls if there were any data irregularities.

b. Sample

For the enrollment study, we used the same schools that were randomly selected for the MELQO student assessment study because that sample is representative of the large regions. See the section on the sample for the student assessment for a full description of the selection process.

c. Analytic plan

We calculated descriptive statistics to examine enrollment and attendance and financial grant and contributions over time. We plotted the data in figures to identify trends and patterns. We explored differences by intervention status and the school's location (or region and district). Finally, we examined students' ages to understand the share of enrollment and attendance for children aged three to seven. Enrollment statistics, such as the pupil-teacher ratio, were also used in supplementary tests of program impacts.

4. Qualitative interviews and focus group discussions

a. Description and design

The purpose of the qualitative portion of the study was to document and track stakeholders' perceptions and ideas about the broad study questions, such as which are the most salient aspects of the FkW model, what improvements are still needed in model components, and what new or persisting challenges undermine the quality of pre-primary education. We investigated and documented the strategies that teachers, schools, communities, and districts implemented to improve the quality of pre-primary and perceptions on how to make those strategies both scalable and sustainable. We also explored stakeholders' views and recommendations on the policy, programmatic, and systemic improvements and adjustments needed to help schools and teachers continuously improve quality and overcome the contextual challenges across Tanzania. Finally, we explored how successful practices and activities can be scaled and sustained in a cost effective manner country-wide. We conducted qualitative interviews with key informants including teachers, both certified and paraprofessionals; head teachers; and the District Academic Office, District Executive Director, Ward Education Officers, and Quality Assurance Officers. Further, we conducted FGDs with parents, community members, and SMCs.

We conducted qualitative interviews and FGDs in 2017 and 2018 for a sample size of 170 transcripts in each year. This data allowed us to track changes in opinions, achievements, and challenges over time. In most cases, we contacted the same informants to best track the evolution of processes, implementation, and perceptions. If a respondent was unavailable we interviewed a replacement at the same location or school.

b. Tools

We developed tools to guide the qualitative activities with key informants. The tools were circulated amongst the Steering Committee members for feedback, and customized for informants at the community, school, ward, and district levels. The tools were translated into Swahili. Next, the data collection team pre-tested the tools and practiced conducting interviews and FGDs with respondents from the corresponding participant group at schools. Subsequently the tools were refined based on input from the partners and lessons learned from the pre-test. Below we describe the focus of each tool:

- The teacher interviews used open-ended questions allowing informants to articulate their perceptions of teacher training and mentoring, instructional methods, implementing the FkW approach and TIE curriculum, the school and classroom learning environment, use of learning materials, classroom management, student learning, enrollment and attendance, school leadership and support, parent partnerships, and community supports.
- The interviews with head teachers focused on perceived changes in teachers' practices and the learning environment, and explored respondents' perceptions of leadership activities to support pre-primary education, support from education officers, capitation grants and the use of funds in pre-primary, implementation of school action plans, and teacher preparedness and instructional practices. We also explored interactions between head teachers and SMCs, WEOs, VEOs, and other local actors, as well as perceptions of the sustainability, scalability, and cost effectiveness of FkW.
- The interviews with the District Academic Office and District Executive Director focused on informants' perceptions of pre-primary education, contextual challenges and overcrowding in

classrooms, and the roles and responsibilities of DAOs and DEOs in supporting education. We investigated the financing of education in general and pre-primary specifically, as well as other potential sources of funding for pre-primary education. We also inquired about the interaction between national, regional, and district offices with regard to education policies, the most salient aspects of the FkW model, and the sustainability, scalability, and cost-effectiveness of FkW. This may include, for example, enforcing age-restriction policies to reduce overcrowding, or implementing child care programs targeted to children under age five who are not ready for pre-primary classrooms.

- The interviews with the WEOs and the Quality Assurance Officers focused on topics related to the oversight and implementation of pre-primary education, including supporting schools, improving the school and classroom learning environment, and efforts and challenges to ensuring quality in pre-primary. We explored informants perceptions' of pre-primary education, the FkW approach, TIE curriculum, recent policy changes, teacher preparedness, teacher training, teacher recruitment and retention, and recommendations on how to improve quality in pre-primary classrooms. We also inquired about informants' interactions or recommendations for SMCs and parent and community engagement.
- The focus groups with parents and community members focused on participants' perceptions of pre-primary education, FkW, the school management and leadership, teachers' instructional practices, the school and classroom environment, and overcrowding and safety. We also asked about parents' and community members' perceptions of their successes and challenges in community engagement and contributions, as well as promising approaches to improving schools that might be replicated. Parents shared their perceptions of the value of the Parent Partnership Program (PPP), how they would assess student's learning, and the role that parents play in ECE.
- The FGDs with SMCs focused on participants' perceptions of pre-primary education, the SMCs' activities related to pre-primary, SMCs' supports and challenges, the school and classroom environment, the FkW model, and school financing and additional sources of support for pre-primary. We also investigated SMC's perceptions of the teacher shortage, the use of paraprofessionals, enforcing national age guidelines for pre-primary students, and parent and community engagement.

c. Sample

We selected the sample for the school and community-based qualitative activities from the schools in the MELQO sample.

- We interviewed 40 certified teachers and 18 paraprofessionals. The sample was split so that we selected eight teachers from intervention schools, eight teachers from control schools, and four teachers from pilot schools in both Moshi and Mwanza. We selected three paraprofessionals from intervention, control, and pilot schools as well. Likewise, we conducted interviews with a sample of 40 head teachers from the same intervention, control and pilot schools as the teacher sample.
- We selected DAOs (n = 2), DEOs (n = 2), WEOS (n = 4), and QAOs (n = 6) within the same district, ward, and villages where the schools for the sampled teachers and head teachers are located. We split the sample across Moshi and Mwanza.
- Finally, we conducted 40 FGDs: 16 with parents, 16 with SMCs, and 12 with community members. We randomly select communities from the sample that we use to conduct the teacher and head teacher interviews. Once communities are selected, we worked with the schools and

local leaders to recruit participants for the FGDs. For the parent FGD, we recruited at least some parents who have participated in the PPPs, however, in control schools, we focused on parents of pre-primary students.

d. Training and field procedures

The field team participated in a training session on qualitative data, which included a thorough review of data collection guides and processes, a description of sampling and recruiting procedures, a discussion and review of high quality transcripts from interviews and FGDs, mock and practice interviews and FGDs, and tool piloting and debriefing. We standardized the data collection approach of the entire team. The training provided mentored time to conduct practice interviews and FGDs and for the team to provide feedback on the tool length and content. Once the team piloted the tools, we discussed the interviews and FGDs to identify areas of success and places to improve.

During field implementation, we closely monitored the entire data collection process. CSR organize and monitor on-the-ground operations and ensured that Mathematica's data quality standards were met. The field researchers followed the sampling guidelines and use the study's tools to conduct their activities. Each interview and FGD was digitally recorded while a note taker took notes by hand or computer to ensure there was no data loss. CSR then used the digital recordings to complete word-for-word transcription of the audio files. The Swahili word files were then translated into English and supervisors randomly selected audio files and transcripts for review to ensure quality.

e. Analytic plan

First, we began the analysis by reading and re-reading the English transcripts. In the initial reading, we identified preliminary classification schemes based on the data. We also identified concepts based on the study's research questions, the qualitative tools, and FkW's program logic. We developed analytic codes and a coding hierarchy that enabled us to explore, sort, and organize key concepts that emerge from the data. Next, we coded the transcripts word by word according to key themes, using NVivo qualitative data analysis software. We reviewed, organized, and analyzed the data based on themes that relate to the program logic and the evaluation questions. We compared responses by respondent type and location to identify similar and disparate themes across respondent groups.

The final analysis involved analyzing the coded data, and then synthesizing and validating responses to extract the key findings related to the various study themes and concepts. We repeated this analytical process until we had mined all of the rich content and nuances from the qualitative data. Once we analyzed each data source, we triangulated findings across the interviews, FGDs, and other relevant data sources and documentation, and integrated findings from the quantitative evaluation components. This process makes it easier to identify new trends and relationships, confirm or validate patterns, and detect discrepancies or disparate findings. In addition, our team participated in a conversation to synthesize the themes by systematically discussing the respondents' perceptions of PPE and FkW and topics relevant to the evaluation questions.

Given that we collected multiple rounds of qualitative data, we presented findings in an iterative manner, building on lessons learned and highlighting cases where challenges have been overcome. We present both summary findings and representative quotes to help the reader understand the themes in more detail. The quotes provide a sense of the stakeholder responses, as well as the varying perspectives of respondents with regard to different themes.

6. Cost analysis

a. Description and design

Next we examined the costs of designing and implementing FkW to help assess the overall merit of the FkW investment.

b. Procedures and analysis

Our primary goal was to capture all of the costs associated with FkW in intervention schools. Given that the sample size for the student assessment was decreased to allow for implementing other study activities, the scope of this activity was also reduced given that it was not reasonable to conduct a full cost effectiveness study with a smaller than optimal sample size to estimate impacts. The new scope was to analyze Dubai Cares' investment in FkW and the allocation of resources by CiC and UNICEF. We collected and examined **administrative data** from Steering Committee partners to calculate and organize all costs associated with the FkW intervention in the schools participating in the evaluation.

D. Ethical Approval

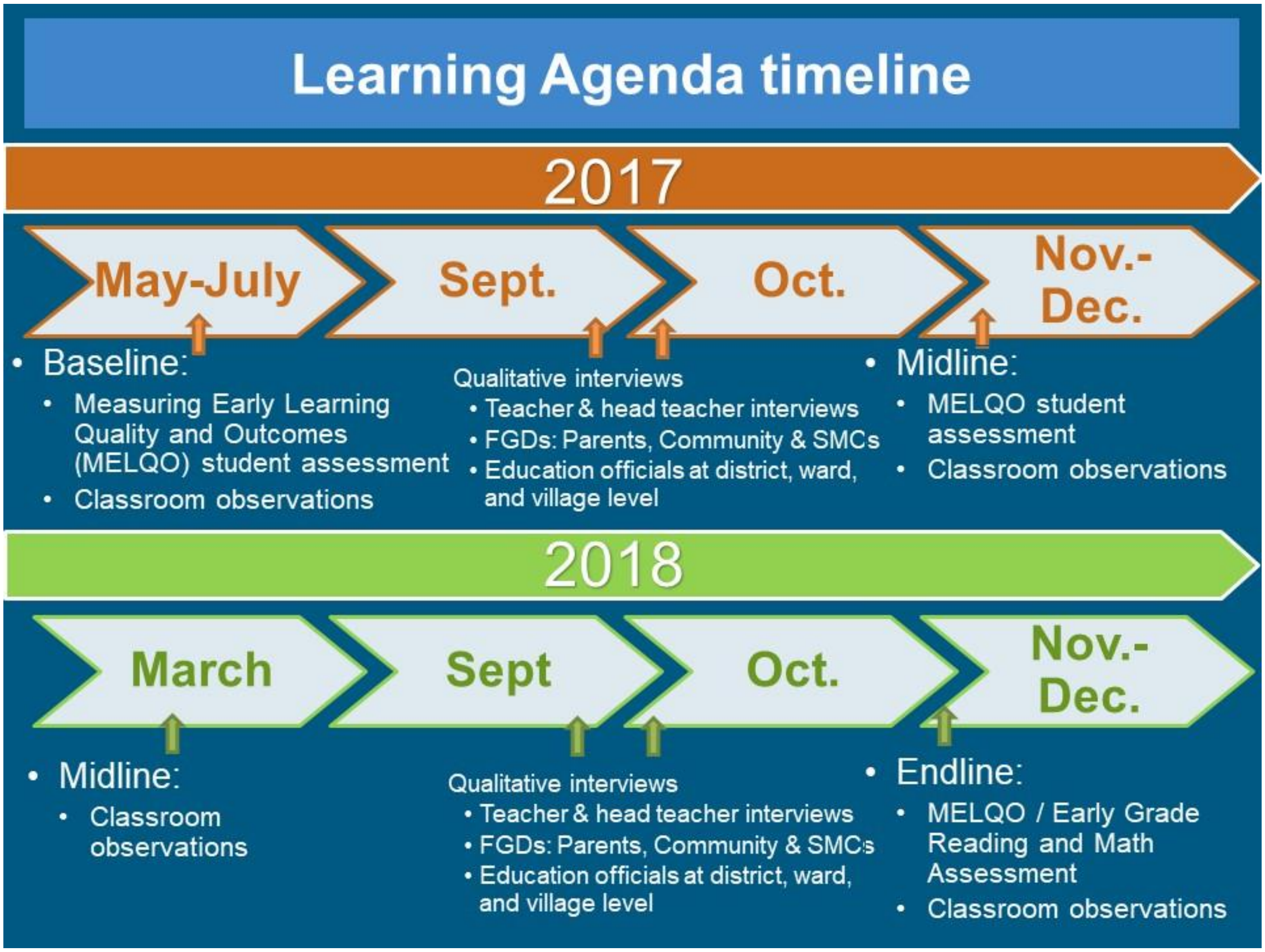
The study was approved by the Tanzania Commission for Science and Technology (COSTECH). The application, including the study design, sampling procedures, instruments, and specific details about how children were consented to participate in the study, was submitted in February 2017 and approved in March 2017.

Before conducting the student assessment, we worked with schools to inform teachers and parents about the assessment. On the day of the assessment, we obtained verbal assent from children before they participated. We ensured that all children knew their participation was voluntary, and they could refuse to participate at any time. We followed all established rules and guidelines for ethical practices in Tanzania. Following the assessments, student data was kept confidential and aggregated to present classroom-level trends.

We have followed all of the guidelines for data dissemination in Tanzania. We presented preliminary results to the President's Office in April 2019 and with government support and participation, we presented results to regional, district, and local stakeholders in Mwanza in July 2019 and Kilimanjaro in August 2019.

E. Timeline for the Learning Agenda

Figure 2. FkW learning agenda: timeline of activities 2017–2018



Learning Agenda timeline

2017

May-July

Sept.

Oct.

Nov.-Dec.

- Baseline:
  - Measuring Early Learning Quality and Outcomes (MELQO) student assessment
  - Classroom observations

Qualitative interviews
- Teacher & head teacher interviews
- FGDs: Parents, Community & SMCs
- Education officials at district, ward, and village level

- Midline:
  - MELQO student assessment
  - Classroom observations

2018

March

Sept

Oct.

Nov.-Dec.

- Midline:
  - Classroom observations

Qualitative interviews
- Teacher & head teacher interviews
- FGDs: Parents, Community & SMCs
- Education officials at district, ward, and village level

- Endline:
  - MELQO / Early Grade Reading and Math Assessment
  - Classroom observations

F. Student results

The results from this study are housed at http://www.fkwlearningagenda.com. This website contains results briefs, power point presentations, tables of data, study tools and instruments, and additional materials. Please contact the Dr. Candace Miller at Mathematica, Inc. for questions or additional results and materials.