Policy Research Working Paper 10482

What Can We Learn from Pre-Primary Quality Assurance Systems?

Evidence from the Arab Republic of Egypt

Caroline Krafft
Samira Nikaein Towfighian
Abbie Raikes
Rebecca Sayre Mojgani

WORLD BANK GROUP
Education Global Practice
June 2023

Abstract

Quality assurance systems have been implemented or are under development in a number of low- and middle-income countries in an effort to observe the quality of education and deploy targeted measures to improve quality. This paper shares lessons learned on the potential ability of quality assurance systems to observe quality and inform action, drawing on data from a pre-primary quality assurance system in the Arab Republic of Egypt. A nationally representative study of kindergarten classrooms was conducted, using a detailed diagnostic research tool administered by independent enumerators from a data collection firm. A subsample of these kindergarten classrooms was randomly assigned to also be observed through a short quality assurance system tool, half of them by independent enumerators, and the other half by the existing cadre of government kindergarten supervisors. The quality assurance system tool was developed for scale and financial sustainability; thus, it could be administered in roughly one-third of the time of the diagnostic tool, at one-third of the cost. Overall, the results illustrate that at the national level, the quality assurance system tool can identify important areas for improvement, and thus inform broad policy actions. Further, the results are consistent whether an independent data collection firm or a government kindergarten supervisor acted as enumerator, suggesting that quality assurance system data collection efforts can be embedded within ministries of education and implemented in a regular and sustainable manner. At the school and teacher level, however, there were several areas where the quality assurance system data were inconsistent with the diagnostic data. This underscores how quality assurance systems are best used as a formative system, a starting point for quality enhancement, and not as a summative system that directly targets, punishes, or rewards specific schools.

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

Produced by the Research Support Team

What Can We Learn from Pre-Primary Quality Assurance Systems? Evidence from the Arab Republic of Egypt

By Caroline Krafft,[1] Samira Nikaein Towfighian[2], Abbie Raikes,[3] and Rebecca Sayre Mojgani[4]

Keywords: Pre-primary; school quality; quality assurance; Egypt
JEL codes: I21, I22, I25, I28

Acknowledgments:

The authors thank the Ministry of Education and Technical Education of Egypt (MOETE) for the support provided throughout the study. Particularly, we thank Nelly El Zayat, Randa Shaheen, Linda Herrera, Kindergarten supervisors and all other MOETE colleagues who supported this work. Rana Yacoub provided excellent research assistance at all stages of the study. The authors are thankful to Amanda Devercelli, Alaka Holla, Juan Baron, Amira Kazem, Adelle Pushparatnam, and Bridget Crumpton for their comments and feedback on earlier versions of the paper. We are also grateful to Andreas Bloom and Marina Wes as well as to the Egypt Country Management for their overall support. Data collection was conducted by Baseera, led by Hanan Girgis. This work was funded by the UK Strategic Partnership for Egypt's Inclusive Growth Trust Fund and the World Bank. All findings, interpretations, and conclusions in this paper are those of the authors and do not necessarily represent the views of the World Bank or the MOETE.

[1] Corresponding Author, Associate Professor of Economics, Department of Economics and Political Science, St. Catherine University, 2004 Randolph Avenue, St. Paul, MN 55105, USA, cgkrafft@stkate.edu ORCID: 0000-0001-6906-9418

[2] Senior Economist and MENA Focal Point for Early Childhood Education. Education Global Practice, The World Bank.

[3] Associate Professor, College of Public Health, University of Nebraska Medical Center and Founder, ECD Measure.

[4] Chief Operations Officer, ECD Measure

1 Introduction

Low- and middle-income countries (LMICs) have made enormous progress in expanding access to education and enrolling children in primary and, increasingly, secondary education. Learning, however, has lagged this growth in schooling (World Bank 2018a). The preparation of children to learn and challenges in early childhood development, including lack of access to quality pre-primary education, further limit learning (Berlinski, Galiani, and Manacorda 2008; Berlinski, Galiani, and Gertler 2009; World Bank 2018a). Recent efforts to expand pre-primary education to enhance learning have underscored the importance of providing high-quality pre-primary education in order to generate developmental gains (Araujo et al. 2016; Bouguen et al. 2018; Blimpo et al. 2019; Wolf et al. 2019).

Quality assurance systems (QAS) have been identified as an important part of promoting quality learning. Two of the three key recommendations of the World Bank (2018a) flagship report on education were to assess learning and act on evidence – actions that can be enabled by a QAS. In line with this recommendation, many countries around the world are developing or strengthening their QAS (Raikes et al. 2020), including the design and implementation of short, feasible tools that can be used to collect data on school- and classroom-level quality indicators. QAS are important particularly for ensuring the quality of early learning, since at this stage of development, there are no national assessments of child development and learning to draw upon when considering policy actions (World Bank 2013; Raikes, Neuman, and Burton 2019). There is a substantial push in conjunction with expanding pre-primary to ensure it is high quality, which has led to a rising number of LMICs working to develop QAS (Raikes, Sayre, and Lima 2021). There has been some research providing advice on the design of QAS and especially standards and monitoring tools for LMICs, including details on case studies and countries' practices (Raikes, Neuman, and Burton 2019; Raikes, Sayre, and Lima 2021).

Yet there has been very little (to the best of the authors' knowledge, none to date) research actually using data from QAS to illustrate their potential – or pitfalls – in supporting quality learning for all in LMICs. Although there is information from high-income countries about their QAS, information from LMICs is limited (Anderson et al. 2017; Raikes, Neuman, and Burton 2019). This is partly a data problem; countries' ministries or agencies responsible for the QAS may release the results of quality assurance efforts, such as school report cards or national reports, but not the microdata nor research validating the QAS itself. Even national reports or databases are relatively rare in LMICs. For instance, only 29% of countries, in a study of early childhood care and education QAS in 14 Sub-Saharan African countries, had their monitoring data reported back to a national database (Raikes, Sayre, and Lima 2021).

This paper investigates the potential ability of QAS to systematically monitor quality and inform action, drawing on data from pre-primary in the Arab Republic of Egypt. The effectiveness of QAS depends on the accuracy and relevance of the data collected through QAS tools used to monitor quality practices within classrooms. Our research questions are:

(1) How consistently do the tools used within a QAS measure quality?
(2) What are potential sources or reasons for any inconsistency?
(3) In light of the degree of measurement error found in measuring quality, how should data from QAS be used to undertake quality enhancement?

A detailed diagnostic research tool and a pilot of the new national QAS tool for Egyptian Kindergartens (KGs) were implemented on a nationally representative sample of schools with KGs, with overlapping data collection. Both government and data collection firm enumerators were used (and randomly assigned to districts). We compare these data sources to assess how consistent results are across the QAS tool and the diagnostic research tool, overall, and for particular items. We specifically measure exact agreement of items and Cohen's kappa coefficient. We do not and cannot know whether the QAS tool or diagnostic research tool data, or neither, are ultimately "right or wrong," but relatively consistent results from different measures are an important prerequisite to data quality and usability. Data were collected on the learning environment and child development outcomes. We examine what school, class, teacher, student, and enumerator characteristics predict greater consistency in logit models for each dimension of quality in Egypt's KG QAS. We then use our findings to simulate whether a variety of different potential quality enhancement strategies would be consistently targeted.

The results, comparing the two tools, are relatively consistent in terms of national averages, suggesting the QAS provides valuable information for national action. On the classroom, school, and district levels, consistency is lower. Consistency varies by the dimension of quality and data collection methods used, with direct measures of child development outcomes or easily observable characteristics, such as class sizes, showing more consistency. There are relatively few relationships between consistency and school or class characteristics, which bodes well for equitable implementation. Moreover, consistency is similar with ministry employees as with professional data collectors, who were randomized across districts. This result is promising for cost-effective national scale up. However, the limited consistency of many items and dimensions means that targeted quality enhancement actions often would target different schools using the two different measures. The results underscore how QAS are best implemented as a formative system, a starting point for quality enhancement, and not as a summative system that directly punishes or rewards specific schools, classes, or teachers.

2 Background

2.1 What we know about education quality assurance systems

QAS (for pre-primary and otherwise) have three main components: (1) quality standards, (2) tools for monitoring whether standards are being met, and (3) quality enhancement actions that follow from the results of the monitoring tool (Raikes, Neuman, and Burton 2019). As a concrete example, consider a pre-primary quality standard that teaching should be play-based. The monitoring tool would then, in this case likely using a classroom observation tool, collect data on whether a play-based pedagogy was being implemented. If the monitoring tool data indicated a teacher was not using play-based pedagogy, a potential quality enhancement action would be providing the teacher training on play-based pedagogy. Although this is the ideal tripartite system, it is important to note that countries may not have all these components in place; for instance in a recent study on pre-primary QAS in Africa, countries usually had at least one component of standards or tools, but only some had all components in place (Raikes, Sayre, and Lima 2021). There is particularly limited comprehensive information globally about quality enhancement components of QAS. While the SABER-ECD database includes 38 countries and looks at monitoring and standards for pre-primary QAS, it does not look at quality enhancement

(World Bank 2023). Moreover, 32 of the 38 countries had latent or emerging quality systems, and no countries had advanced quality standards or compliance with standards (World Bank 2023).

2.1.1 Quality standards

Defining quality education and specifically quality pre-primary education is challenging. What defines a quality context depends on how teaching practices and learning environments impact development, what dimensions of development are prioritized, and also national educational goals. Quality standards are often designed to index both the basic requirements for safety as well as aspirations for pedagogy within classrooms, and may also set requirements for teacher training, class sizes, and other aspects of quality (Raikes, Sayre, and Lima 2021; Bendini and Devercelli 2022).

2.1.2 Tools for monitoring

Monitoring tools for QAS need to both measure the underlying standards (be valid) and provide consistent results (be reliable). Tools need to be administered periodically and at scale in order to regularly monitor quality for an entire country. There are a variety of different research tools designed to measure early childhood environments and early learning (Fernald et al. 2017). There is a rich literature examining the psychometric properties of various existing research tools (but not usually monitoring tools) to assess early childhood development (McCoy et al. 2018; Raikes et al. 2019) as well as studying the functioning of various classroom observation tools (Molina et al. 2018; Wolf et al. 2018) (but again, not necessarily those used in QAS). Such tools are primarily intended for research, requiring highly-trained observers and not designed to measure country-specific standards, nor to provide consistent results when administered by government supervisors (with limited training) at scale (Fernald et al. 2017). These measures are also not intended for high-stakes decision-making, for example, whether classrooms should receive praise for exceptional performance or should be targeted for improvement. This has led many LMICs to adapt these tools or develop nimbler tools for monitoring as part of a QAS (Raikes, Sayre, and Lima 2021).

To the best of our knowledge, there has not been research using data from QAS tools from LMICs to assess their consistency or implications of inconsistency for targeting quality enhancement actions. In high-income contexts, particularly the United States, there has been some research on early-childhood QAS, typically named Quality Rating and Improvement Systems (QRIS). Research on QRIS highlights a number of measurement issues that are likely to be pertinent for QAS as well. For instance, one study of Minnesota’s QRIS showed item-level correlations that were weak and variable correlations between different categories that went into overall scores (Tout et al. 2011). Validation of QRIS ratings against other measures of quality generally find positive correlations (Zellman et al. 2008; Elicker et al. 2011; Tout et al. 2011). However, QRIS measures are not necessarily reliable or sizable predictors of children’s outcomes (Zellman et al. 2008; Elicker et al. 2011; Tout et al. 2011; Keys et al. 2013). These findings demonstrate some of the challenges of developing and implementing QAS tools.

2.1.3 Quality enhancement actions based on QAS tool data

A key question for QAS tools is how the data will be used. With constrained education budgets, ideally, the results of the monitoring tool can inform LMICs on how to tailor and target quality enhancement actions. Examples of quality enhancement actions include deployment of resources (such as furniture, repairs, materials, or funds) to schools that are in need, teacher support (training or coaching responding to topics and teachers that are struggling), or incentives (including information, or targeted rewards, recognition, or accountability). Although there is not much research on the effectiveness of these quality enhancement actions as different components of QAS, there is a large body of research on what works to promote learning in LMICs (Glewwe et al. 2013; Krishnaratne, White, and Carpenter 2013; McEwan 2015; Evans and Popova 2016; Ganimian and Murnane 2016; Glewwe and Muralidharan 2016; Conn 2017; Evans and Mendez Acosta 2021).

Additional resources can improve learning, but the impact of local management of funds is mixed and depends on the capacity of management committees (Blimpo and Evans 2011; Glewwe and Maïga 2011; Pradhan et al. 2014; Santibanez, Abreu-Lastra, and Donoghue 2014). Teaching and learning materials, such as slates for students or scripts for teachers, can lead to quality improvements, and act as important complements to training and coaching (Glewwe et al. 2004; Rolla San Francisco et al. 2006; Glewwe, Kremer, and Moulin 2009; Cristia et al. 2012; Piper et al. 2018). Training as it is often realized in LMICs (one-off, centralized, in a cascade model) tends not to be effective (Desimone et al. 2003; Yoon et al. 2007; Sayre, Raikes, and Devercelli 2018; Wolf 2018; Blimpo et al. 2019). Practice, feedback, and longer (but more focused) training and coaching tend to be more effective (Pallante and Kim 2013; Westbrook et al. 2013; Reinke et al. 2014; Fleisch et al. 2016; Kotze, Fleisch, and Taylor 2019; Popova et al. 2022).

In terms of incentives, information, and accountability, school report cards or quality ratings (which could potentially report QAS results) can cost-effectively improve education (Andrabi, Das, and Khwaja 2017; Bassok, Dee, and Latham 2019), but if implementation is weak they will have no effects (Aturupane et al. 2014). The global evidence on pay-for-performance for primary teachers is mixed (Glewwe, Ilias, and Kremer 2010; Muralidharan and Sundararaman 2011; Neal 2011; Goodman and Turner 2013). Although there is this rich literature on what works for education, how these policies would play out with a QAS has not previously been investigated.

2.2 Kindergartens in Egypt

Children aged 4-6 in Egypt are eligible for KG, which is not compulsory. Children who enroll in KG can do so at KG1 or KG2 grades. At age six, children are eligible for primary school. Kindergartens are overseen by the Ministry of Education and Technical Education (MoETE) in Egypt, which provides public KG classes in public primary schools. Most KG enrollment is in public KGs, but a substantial private KG sector, primarily enrolling wealthier families, exists (El-Kogali and Krafft 2015). In the 2010s, enrollment rates in pre-primary education in Egypt were around 28% (World Bank 2022).

In its “Vision 2030” national plan, MoETE set a pre-primary education enrollment target of 80% by 2030—close to three times its 2015 enrollment (Ministry of Planning and Economic Development 2015). While aiming towards an unprecedented expansion in access to pre-primary education, MoETE also emphasized the need to improve quality in service provision (Ministry of Planning and Economic Development 2018; Moustafa et al. 2022). These parallel goals led MoETE in 2019 to start designing a QAS for pre-primary education that could regularly monitor and assure quality in a sustainable manner and at an increasingly large scale (World Bank 2018b).

3 Methods

3.1 Measures

We provide two measures of the consistency of both specific items (items are described below in the data section) and QAS levels (school, classroom, dimension, and sub-dimension). First, we present the percentage of items or categories with exact agreement (consistent responses), the “agreement coefficient” (Gwet 2014) across the diagnostic research tool and QAS tool measures. Exact agreement is simple to interpret but can happen by chance and also can be driven by the underlying distribution of an item. Cohen’s kappa coefficient accounts for the probability of chance agreement, p_e, based on the probabilities observed in the data, and uses this and the actual agreement, p_a to calculate (Gwet 2014):

$$\kappa = \frac{p_a - p_e}{1 - p_e}$$

Kappa is designed to work with categorical data (most of the measures are binary and levels of performance on sub-dimensions and dimensions are ordered categories). We use the standard Landis-Koch benchmark scale (Landis and Koch 1977) to classify strength of agreement based on the kappa. Less than 0.00 is poor; 0.00-0.20 is slight agreement; 0.21-0.40 is fair agreement, 0.41-0.60 is moderate agreement; 0.61-0.80 is substantial agreement, and 0.81-1.00 is almost perfect agreement. While we would not realistically expect 100% exact agreement or almost perfect agreement on all items, items with substantial or almost perfect agreement will lead to better measurement and ultimately a better functioning QAS.

3.2 Models

We estimate a series of multivariate logit models for an outcome of consistency on each dimension of the QAS at the classroom level. Denote as $y_{i,j}$ the consistency (0=inconsistent, 1=consistent) for dimension j in classroom i. An observation in this case is effectively collapsed from a pair of observations (QAS tool and diagnostic research tool). We relate this outcome to the covariates, denoted $X_{i,k}$, for k different covariates. Our logit model for the probability of consistency $p(y_{i,j}=1| X_{i,k})$ is thus:

$$ln\left(\frac{p(y_{i,j} = 1|X_{i,k})}{1 - p(y_{i,j} = 1|X_{i,k})}\right) = \beta_0 + \sum_{1}^{k} \beta_{i,k,j} X_{i,k} + \varepsilon_{i,j}$$

We present exponentiated coefficients (odds ratios) from these models, along with standard errors clustered on the enumerator level.

3.3 Simulating the targeting performance of the QAS

To understand what we can learn from QAS, we simulate the performance of the QAS in terms of targeting different potential quality enhancement actions that could follow from the monitoring tool data of the QAS. We refer to these analyses simulating the targeting performance of the QAS for quality enhancement as "simulations" for short. Specifically, we compare who would be the targets of different policy actions using the QAS tool and diagnostic research tool data. We examine:

- Differences in the top 10% and top 25% of districts, schools, and classes/teachers identified with the QAS overall, which might be used for recognition, pay, or promotion
- Differences in the bottom 10% and top 25% of districts, schools, and classes/teachers per the QAS overall, which might be used for shutdown, termination, or other sanctioning decisions
- Differences in targeting repairs (whether the bottom 10% and bottom 25% of districts, schools, or classes with the greatest number of safety hazards is consistent)
- Differences in targeting materials (whether the same districts, schools, or classes would be targeted for supplemental grants and training for materials, based on not having or using any materials)
- Differences in targeting pedagogy training or coaching (based on the poorest performing level in terms of pedagogy) for districts, schools, and classes
- Differences in supplemental reading tutors, based on the poorest-performing 10% and poorest-performing 25% on children's letter recognition for districts, schools, and classes

Differences are parametrized as the percentage of the simulation targets in the QAS tool data that are identified the same way in the diagnostic research tool data. These different simulations using the QAS levels, the level of a particular QAS dimension, and specific items, illustrate how these different approaches to using QAS data to inform action might perform.

We examine these questions at the district, school, and class levels for a variety of reasons. First, policies might operate on these different levels, depending on the design of the quality enhancement system and administration. Second, measurement error may be reduced when looking at aggregates, particularly if outcomes in a school, class, or district are highly correlated. We examine, for the outcomes that target a certain ranked (top or bottom) percentage, both 10% and 25% targeting. The narrowness or coarseness of targeting may affect consistency. These simulations do not incorporate covariates, or use the multivariate models, but rather describe the average, national consistency of the QAS system as a data source for quality enhancement action.

4 Data

4.1 Data collection tools, training, and fieldwork

The data collection in Egyptian Kindergartens was designed to further two goals: (1) to do the first ever in-depth diagnosis of the quality of teaching and learning in Kindergartens, and (2) to pilot and validate a nimble monitoring tool that Egypt's Ministry of Education can afford to administer periodically as part of its QAS. There were thus two overlapping data collection efforts, what we refer to as "the diagnostic research tool" and "the QAS tool."

4.1.1 The diagnostic research tool

The diagnostic research tool used as its starting point the Measuring Early Learning Quality Outcomes (MELQO) tools (UNESCO 2017). The MELQO tools were developed in order to measure child development and quality of early childhood education in LMICs (Raikes et al. 2019).[5] These tools were locally adapted to the Egyptian context, including to reflect the Egyptian KG curriculum. The curriculum had recently been updated as part of the "Education 2.0" reforms (Moustafa et al. 2022). There are two main parts to the MELQO tools: (1) the Measure of Early Development and Learning (MODEL) measures the development of children aged 3-6, and (2) the Measure of Early Learning Environments (MELE) measures learning environments and their quality.

For this study, we use MODEL data collected through a child direct assessment, a parent report of child development (which includes family background), and a teacher report of child development. We use MELE data collected through classroom observation, a teacher interview, and the school director interview.

4.1.2 The QAS tool

The starting point for Egypt's KG QAS tool was the Brief Early Childhood Quality Inventory (BEQI).[6] The BEQI was built on the MELQO tools, but was designed to be a simpler tool than MELQO, integrated into monitoring systems or used for formative assessment (ECD Measure 2022). As an example of simplification, items that were measured on a four-point scale in MELQO were typically transformed to yes/no for the BEQI. Some items, however, were identical across the two tools. The BEQI was adapted to the Egyptian context, including Egypt's recently adopted KG quality standards (see Appendix 1), as discussed in more detail below. A summary of the features of the diagnostic research tool and QAS tool is provided in Appendix 2.

4.1.3 Tool adaptation

The BEQI and MELQO tools were adapted to the Egyptian context and curriculum in collaboration with the MoETE, Kindergarten teachers, and Kindergarten supervisors. Both tools were adapted in a May 2019 workshop. For data collection, tools were programmed into Android

[5] The MELQO tools have been used and validated in other LMIC contexts (Raikes et al. 2020).

[6] The BEQI has been used in other country contexts as well (Raikes et al. 2023).

tablets using ODK-X software (Brunette et al. 2017). Pre-piloting of the instruments took place in Egypt in ten classrooms and with 30 children and adjustments to the tools were made based on the pre-pilot as well as feedback in the subsequent training.

4.1.4 Training

The international experts (members of the author team) trained the master trainers, who included MoETE officials, KG supervisors, and Egyptian academic experts, in January 2020. There were different training programs for the QAS tool and the diagnostic research tool. Training of diagnostic research tool enumerators took place over 10 days, while training of QAS tool enumerators took place over 3 days. The shorter training was designed to both reflect the shorter tool and what would be feasible and affordable for MoETE for training QAS tool enumerators at scale. Training started in late February 2020 and included piloting in schools for both groups. Enumerators were graduates of faculties of Kindergarten education or child psychology, or Kindergarten teachers or supervisors. For the QAS tool, the plan was for half of the QAS tool enumerators to be hired by the data collection firm and for half to be current MoETE employees (similar to those who would eventually implement and scale up the QAS). Before moving to data collection, all supervisors and enumerators were required to pass a written quiz regarding tool content and training procedures and a classroom video quiz with at least 80% agreement with master codes.

4.1.5 Data collection

Data collection was planned for mid-March 2020. On the day data collection was scheduled to start, schools were closed due to COVID-19. Data collection was thus delayed for a year due to COVID-19. After schools reopened in October 2021, a repeat training was held for enumerators.[7] Data collection in schools took place from November 6, 2021, to December 8, 2021. Parents were interviewed over the phone through December 15, 2021.

For the diagnostic research tool, which was a lengthier and more detailed tool, enumerators in the schools were specialized into one of three roles: supervisor (who undertook the director interview and logistics), classroom observer (who also did the teacher interview) and child direct assessor (who also did the teacher report of child development). QAS tool enumerators entered the schools with the diagnostic research tool teams and started classrooms with the classroom observer and child enumerator to have the same (random) sample of children. Since they also did the child direct assessment and teacher child report, they may have observed a different portion of the class and would have been assessing children at different times than the child direct assessor.

4.2 Sample

The sampling frame was Egypt's Education Management Information System (EMIS) database from 2018-19. We stratified the sample by school type (public versus private), region,[8] and

[7] Enumerators were mostly new, so the full training was repeated.

[8] Regions were divided into: Urban Governorates, Lower Egypt, and Upper Egypt. Frontier Governorates, which have only a small fraction of the population, were not included.

community poverty status. Within each stratum, a random sample totaling 46 districts was drawn.[9] Five schools were randomly selected within each district.[10] The resulting sample was nationally representative of Egyptian schools with KGs.

A total of 214 schools were sampled for the diagnostic research tool.[11] A random sub-sample of 115 of those schools were selected for the QAS tool data collection effort, with 55 initially randomly assigned to data collection firm enumerators and 60 assigned to MoETE supervisors. Due to limited availability of MoETE supervisors, 18 schools were reassigned to the data collection firm enumerators and five were not collected (due to their location and insufficient availability of data collection firm enumerators in that location).

Data were collected for up to three KG1 and three KG2 classes per school (randomly selected if more than three). A total of 333 classrooms were sampled across schools. Since there was often more than one teacher per class, this led to a sample of 434 teachers consenting to be interviewed and responding to both tools.[12] A random sample of four children per classroom was selected (1,332 children therefore should have been sampled) and 1,169 children consented and responded to both tools.[13]

The data collection firm tried up to three times to reach parents, based on phone numbers provided by the school. For the parent data, there was substantial non-response (primarily that parents did not pick up, but some refusal when reached). Ultimately, 625 parents were reached and consented to be interviewed. Weights are used in all our analyses. The weights account for the original sampling design and non-response.[14]

4.3 Outcomes

Our primary outcomes are the consistency of items or dimensions across the two tools, which are essentially an issue of measurement error. Reliability (internal consistency or test-retest) is a key issue driving potential measurement error (Bound, Brown, and Mathiowetz 2001). Some of the measures we observe are likely to be inherently variable and have modest internal consistency when measured at different times; for example, children may be worse at direct assessment items prior to snack or recess and perform better after snack or recess. Teachers may use different teaching strategies for different lessons, such that the results of observation tools would vary for different days and time periods. The CREDI tools, which assess ECD through parent reports,

[9] Districts were randomly selected probability proportional to size (based on the number of schools), with replacement. Districts were drawn from within regions and based on the poverty status within a region (33% poor schools as cutoff).

[10] If there are fewer than five schools within a district and strata, all schools were used (one to four).

[11] Seven of the originally selected schools were unavailable (closed, in renovations, etc.) and random replacement schools, as much as possible from the same strata, were used.

[12] The teacher consent rate was 87% for the diagnostic tool interview and 94% for the (shorter) QAS tool interview.

[13] This is an overall response rate of 88%. Non-response was about 8% for each of the two tools and 12% overall, as children tended to refuse both more often if they refused at least one.

[14] The weights account for the original sample design on the district and school levels, the sampling of classes (for class and teacher outcomes), and the random sampling of students (for student outcomes). Weights account for non-response and the number of observations that should have been included, for instance, the number of parents per class there should have been.

checked test-retest one week apart, had a kappa (measure of agreement) of 0.62 (McCoy et al. 2018).

Inter-rater reliability is also an issue; even if they were observing the *exact* same phenomena, enumerators might provide different measures and responses (Gwet 2014). Enumerators, during training, were required to achieve scores of at least 80% on activities and quizzes in an effort to support inter-rater reliability. As a point of reference, the *Teach* classroom observation tool, designed for LMICs, in a validation study where there were two enumerators engaging in identical observations in the classroom, had an ICC of 0.75. Across categories exact agreement on items between raters ranged from 54% to 79% (Molina et al. 2018).

We focus on an outcome of consistency, since our data collection setup does not allow us to distinguish these different drivers of measurement error. This focus also better reflects how QAS are actually implemented: with different raters and at different times. We look at consistency both for individual monitoring tool items and for different dimensions and subdimensions of the QAS.

In terms of individual items, the QAS tool includes:

Basic classroom and teacher information:

- Class size
- Number of teachers
- Primary teacher highest education level
- Primary teacher specialization
- Primary teacher year started teaching
- In-service training in the past 12 months
- Topics of in-service training

Materials

- Writing utensils (none present; present but children do not use; present some children use; present all children use)
- Manipulatives (none present; present but children do not use; present some children use; present all children use)
- Classroom management tools (none present; present but teachers do not use; present teachers use)

Pedagogy (yes/no)

- Teacher engages in individual instruction
- Teachers and children have back-and-forth discussion/dialogue
- Teacher asks open-ended questions
- Teachers use strategies of Egypt's education 2.0 curriculum
- Teacher connects lessons to real life
- Teacher is mostly positive (warm, responsive)
- Teacher redirects misbehavior

Facilities (yes/no)

- Child-safe and sized seating
- Child-safe and sized desks/tables
- Activity space for all children to get up/engage
- Safe activity spaces outside the room (e.g. gym, playground)
- Size-appropriate, sex-segregated, sanitary toilet facilities
- Soap and running water
- Children wash with soap and water
- Sanitary drinking water

Hazards (yes/no)

- Broken or uneven floors
- Chairs or tables are broken
- There is a leak in the ceiling or holes in the ceiling
- Broken windows or doors
- Natural light is not enough
- Ventilation is not sufficient
- Rocky fields with open trash or pits
- There is no wall around the school building
- The school is close to major roads
- Other conditions that may cause injury to children

Teacher report of child's life (socio-emotional) skills (yes/no)

- Keeps working until finished
- Follows instructions
- Takes into consideration other people's feelings

Child direct assessment

- Letter names: Arabic (repeated for eight letters) (correct/incorrect [includes don't know, no response])
- Letter names: English (repeated for eight letters) (correct/incorrect [includes don't know, no response])
- Name writing (correct/incorrect [includes don't know, no response])
- Verbal counting (highest number)
- Naming shapes (circle, triangle, rectangle) (correct/incorrect [includes don't know, no response])
- Name of kindergarten (correct/incorrect [includes don't know, no response])
- Name of country (correct/incorrect [includes don't know, no response])
- Where fish live (correct/incorrect [includes don't know, no response])
- Can point to flag of Egypt (correct/incorrect [includes don't know, no response])
- Can point to picture of cloud (correct/incorrect [includes don't know, no response])

Almost all of these items are also included in the diagnostic research tool but may have more complex question designs (e.g. 4-point scales rather than yes/no).

These items are grouped into three dimensions and seven sub-dimensions in the QAS:

- Infrastructure and materials (sub-dimensions: Infrastructure; Materials)
- Teaching and Pedagogy 2.0 (sub-dimensions: Teacher's supports; Pedagogy 2.0)
- Child learning and development (sub-dimensions: life skills; foundational skills; multidisciplinary)

The KG quality standards include specific details on how dimensions, sub-dimensions, and classes and schools overall are classified into four levels of performance.[15] For example, for materials, a class is considered (1) below minimum if missing or not using writing and drawing tools; (2) minimum if using writing and drawing tools; (3) developing if using writing and drawing tools and classroom management tools; and (4) achieved if using writing and drawing tools, classroom management tools, and manipulatives. The full standards, including classification into levels of performance, are presented in Appendix 1.

There are also specific procedures related to aggregating scores from the class to school level. To translate classroom scores to school-level scores, the average (1=below minimum, 2=minimum, 3=developing, 4=achieved) across classrooms is taken, and scores then can be rounded to the school level for each sub-dimension. To translate across sub-dimensions to an overall score, levels of performance on each dimension are averaged, equally weighted, and rounded to a school-level score.

4.4 Covariates

Consistency of responses can depend on differences in natural variability, but also can depend on the characteristics of the observer and the observed. The models (discussed above) include controls for MoETE enumerators versus data collection firm enumerators for the QAS tool. We are particularly interested in this dimension of consistency because, while data collection firm enumerators may offer an expert and independent perspective, such an approach is not financially sustainable for governments to implement on a periodic basis. The models also control for whether it is a public or private school, the region, and whether it is a high-poverty school (poverty rate 50% or more from the national poverty map).

From teacher and classroom observations, based on the (more detailed) diagnostic research tool data, our controls include class size, class level, having a second teacher, primary teacher age, primary teacher years teaching, primary teacher years taught at this school, and primary teacher professional status. From classroom observations, we also control for the differences between QAS tool and diagnostic research tool in the start time (in [fractional] days), to capture whether the same period is observed. From child data, our controls include, based on the (more detailed) diagnostic research tool data, child sex, child age (in months), and class level.

We estimate some models with additional controls, specifically socioeconomic status (SES) controls. For school and classroom level outcomes, we take the average asset score, the percentage of mothers (and likewise fathers) with secondary education, and similarly for higher

[15] The dimensions do not simply average indicators within the dimension.

education, the percentage of mothers who work, and the percentage of fathers who have professional/managerial jobs as well as the percentage who work in sales/service.

5 Results

5.1 Measures of consistency

In Table 1 we present mean proportions from the QAS tool and diagnostic research tool for each indicator. We also present mean class level and school level categorical and overall ratings, ranging from 1-4 (below minimum to achieved). An important initial finding is that, on a national level, the tools are generally showing similar results about quality for items, levels, and overall, based on similar means. Items with the exact same question in both tools tend to be more similar, as well as items that are easy to observe, such as inadequate light (10.9% in both tools), and class size below 36 (57.5% in the QAS tool, 56.7% in the diagnostic research tool). More subjective measures as well as measures based on different questions vary more, such as pedagogy or some infrastructure questions, for instance the different questions for adequate outdoor space (66.7% meet standard using data from the QAS tool and 80.1% in the diagnostic research tool).

Questions about children's life skills do not yield similar means, but this may be because a different scale of responses was used for the two tools, even with the same questions. Foundational and multi-disciplinary skill questions, which are identical (questions and responses), do yield similar means. Differences in pedagogy and materials may also be related to observation time. For instance, the QAS tool captures a much lower level of having and using writing utensils (63.9% meet standard versus 84.6% for the diagnostic research tool data) and classroom management tools (41.6% meet standard versus 77.1% for diagnostic research tool data).

Turning now to the measures of agreement; it is important to keep in mind that exact agreement is often high by chance when items are rare or universal (e.g. desks, 90.2% exact agreement, as desks are 94.8% at standard in the QAS tool and 91.9% in the diagnostic research tool). We therefore focus our discussion of the agreement results on the kappa.

Kappa is substantial (0.6 to <0.8) or almost perfect (0.8-1.00) for a limited number of specific items and only among overall and the sub-dimensions for teacher supports at the school level (0.625) and classroom level (0.636). The only items with almost perfect agreement are the class size and the student-teacher ratio. The items with substantial agreement were other hazards, being trained on the curriculum, and the three foundational skills (letter recognition, name writing, and counting to 10).

On the other end of the spectrum were the items with poor kappa (indoor space, kappa=-0.043), and slight (0 to <.0.2) strength of agreement (writing utensils and manipulatives; individualized teaching, correcting misbehavior, discussion, open-ended, relevant pedagogy, and hardworking children). These items were all based on the classroom observation except for children being hardworking, and also would have been potentially time variable (unlike hazards) and subjective. There was fair (0.2 to <0.4) strength of agreement for desks, seating, washing, light, ventilation, classroom management tools, the curriculum, and the life skills of instructions and being considerate. These are generally more readily observable and less subjective items, but also ones where different questions or scales could lead to different measurements.

The moderate strength agreement items (0.4 to <0.6) were outdoor space, toilets, soap, a KG degree, continuous professional development, positive teaching, and all the multi-disciplinary items. The multi-disciplinary items have some potential for test-retest variation in children's responses (but the better-performing foundational skills did as well), whereas the other items may be more readily measurable.

In terms of the class level and school level strength of agreement, interestingly, although categories and overall were weak, classroom level measures performed slightly better than school level averages in terms of kappa. Overall consistency was slight (kappa=0.174 on the class and 0.013 on the school level). Because most schools were at the minimum level of overall quality, exact agreement was higher (82.2% on the class level and 86.7% on the school level).

Table 1. Proportions from QAS tool, diagnostic research tool, exact agreement, and kappa

-	**Mean QAS tool**	**Mean diagnostic research tool**	**Exact agreement**	**Kappa**	**N (Obs.)**
Infrastructure					
Desk	0.948	0.919	0.902	0.211	332
Seat	0.975	0.919	0.925	0.260	332
Indoor space	0.708	0.588	0.517	-0.043	332
Outdoor space	0.667	0.801	0.761	0.402	332
Toilets	0.667	0.525	0.766	0.524	332
Soap	0.493	0.664	0.758	0.518	332
Washing	0.300	0.121	0.729	0.221	332
Inadequate light	0.109	0.109	0.873	0.347	332
Inadequate ventilation	0.094	0.147	0.855	0.320	332
Other hazards	0.588	0.646	0.816	0.613	332
Class size	0.575	0.567	0.980	0.958	333
Student-teacher ratio	0.203	0.215	0.982	0.947	333
Materials					
Writing utensils	0.639	0.846	0.664	0.169	332
Manipulatives	0.250	0.253	0.656	0.086	332
Classroom management	0.416	0.771	0.564	0.200	332
Teacher's supports					
KG degree	0.748	0.628	0.760	0.451	434
Trained on ed. 2.0	0.257	0.279	0.879	0.692	434
Cont. Prof. dev.	0.132	0.108	0.884	0.453	434
Pedagogy 2.0					
Curriculum	0.719	0.906	0.748	0.217	332
Individualized	0.863	0.894	0.786	0.001	332

-	**Mean QAS tool**	**Mean diagnostic research tool**	**Exact agreement**	**Kappa**	**N (Obs.)**
Positive	0.941	0.955	0.943	0.428	332
Correct misbehav.	0.575	0.748	0.584	0.100	332
Discussion	0.787	0.656	0.607	0.042	332
Open-ended	0.404	0.656	0.508	0.073	332
Relevant	0.445	0.656	0.525	0.082	332
Life skills					
Hardworking	0.773	0.939	0.785	0.174	1163
Instructions	0.815	0.940	0.837	0.268	1163
Considerate	0.849	0.929	0.844	0.223	1163
Foundational skills					
Letters	0.210	0.288	0.873	0.663	1169
Write name	0.546	0.604	0.851	0.696	1167
Count to 10	0.897	0.887	0.937	0.672	1167
Multidisciplinary					
KG name	0.470	0.467	0.776	0.550	1167
Country	0.495	0.500	0.795	0.591	1167
Fish	0.941	0.924	0.936	0.490	1167
Flag	0.908	0.901	0.903	0.438	1167
Class level					
Infrastructure	1.067	1.051	0.981	0.552	332
Materials	2.135	2.749	0.311	0.121	332
Teacher supports	1.988	1.883	0.777	0.636	335
Pedagogy	1.482	1.793	0.477	0.112	332
Life skills	3.798	3.960	0.830	0.111	311
Foundational skills	1.033	1.103	0.933	0.434	311
Multidisciplinary	1.598	1.547	0.678	0.386	311
Overall	1.939	2.042	0.822	0.174	336
School level					
Infrastructure	1.098	1.058	0.950	0.292	109
Materials	2.213	2.775	0.331	0.116	109
Teacher supports	2.080	2.053	0.801	0.625	110
Pedagogy	1.400	1.773	0.506	0.202	109
Life skills	3.841	3.959	0.851	0.072	105
Foundational skills	1.021	1.120	0.901	0.267	105
Multidisciplinary	1.613	1.477	0.624	0.329	105
Overall	1.967	2.038	0.867	0.013	110

Source: Authors’ calculations

5.2 Models of consistency

We turn now to multivariate models of QAS level consistency for each dimension of the QAS for classrooms (Table 2). Given high rates of exact agreement for infrastructure and overall, we lack the variation to examine these particular outcomes in a multivariate framework. For each dimension, we present a model with school level controls and then a second specification adding parent SES. We also discuss the pseudo-R-squared, as a measure of how much variability in consistency is explained by the model as a whole. The overall limited patterns of consistency by covariates are promising in terms of limited bias in the measures. This is reflected in the pseudo-R-squared as well, which range from 0.062 to 0.174 (out of a potential 0-1, with zero indicating the covariates have less explanatory power and 1 indicating more explanatory power) in models without parent SES, and from 0.113 to 0.214 in models with parent SES.

Notably, community poverty and parent SES rarely act as statistically significant predictors of consistency.[16] Private schools tend to have significantly more consistent results, specifically for classroom materials, children's life skills, and children's foundational skills. This is possibly in part a ceiling effect, since children from higher SES backgrounds are more likely to attend private schools (Krafft, Elbadawy, and Sieverding 2019). There are some regional differences, with Upper Egypt and Lower Egypt (for different metrics) being significantly more consistent than the urban governorates. Interestingly and importantly, there are not significant results by data collector, suggesting similar consistency for MoETE government employees (who will be implementing the system at scale) as for the enumerators hired by the data collection firm.

Class size is not a significant predictor of consistency, but for the multidisciplinary dimension KG 2 classes have significantly less consistency than KG 1. A second teacher predicts significantly lower consistency for teachers' assessment of children's life skills, which may relate to teachers in larger classes with two teachers knowing individual children less well. There are no significant differences by teacher age or years teaching, but teachers who taught longer in this school rate children more consistently on life skills; they may have known KG 2 students during KG 1 or known siblings of the students and have a better sense of skills. In terms of teacher status, there is only one significant result, less consistency in ratings on materials for senior teachers compared to teachers in one model. Differences in date do predict significantly less consistency, particularly for pedagogy but also teacher supports in one model. The result in terms of pedagogy emphasizes that enumerators may genuinely observe variable practices on different dates, which is likely to be even more of an issue implementing at scale.

[16] While community and parental SES do not predict consistency, they do predict early childhood outcomes (Krafft et al. 2023).

Table 2. Logit models for outcome of QAS level consistency, across QAS dimensions

	Materials	Materials	Teacher supports	Teacher supports	Pedagogy	Pedagogy	Life skills	Life skills	Foundational skills	Foundational skills	Multi-disciplinary	Multi-disciplinary
Community poverty (non-poor omit.)												
Poor	0.764	0.780	0.939	0.632	1.178	1.508	1.470	1.503	1.408	1.400	1.099	1.212
	(0.438)	(0.394)	(0.485)	(0.281)	(0.761)	(0.906)	(0.876)	(0.941)	(1.039)	(1.100)	(0.618)	(0.629)
School type (public omit.)												
Private	4.301*	15.734*	4.751		1.483	1.448	4.451*	1.291	1.840	16.450**	0.901	1.646
	(2.838)	(18.219)	(4.057)		(0.951)	(1.065)	(3.001)	(1.636)	(1.456)	(15.839)	(0.536)	(0.843)
Region (urban govs. omit.)												
Upper Egypt	1.165	1.393	0.332	0.331	2.851	2.846*	2.002	2.451	3.483*	3.873*	0.517	0.548
	(0.456)	(0.580)	(0.239)	(0.201)	(1.589)	(1.230)	(1.054)	(1.290)	(2.084)	(2.116)	(0.282)	(0.279)
Lower Egypt	2.432	2.209	8.491*	19.082**	1.289	0.926	1.108	1.165	0.666	0.588	1.688	1.537
	(1.634)	(1.262)	(7.598)	(19.392)	(0.877)	(0.595)	(0.732)	(1.003)	(0.475)	(0.396)	(1.001)	(0.800)
Data collector (firm omit.)												
Ministry	0.798	0.373	2.362	2.585	0.564	0.571	0.716	0.587	0.872	0.474	0.603	0.601
	(0.437)	(0.238)	(1.212)	(1.295)	(0.276)	(0.264)	(0.326)	(0.365)	(0.620)	(0.219)	(0.180)	(0.227)
Class size	0.993	1.007	0.978	0.988	1.005	1.021	1.018	1.014	1.011	1.015	0.978	0.985
	(0.015)	(0.018)	(0.013)	(0.019)	(0.009)	(0.012)	(0.010)	(0.010)	(0.026)	(0.020)	(0.016)	(0.016)
KG grade (one omit.)												
Grade 2	1.275	1.150	1.045	1.017	0.864	0.887	1.362	1.256	0.437	0.597	0.320***	0.338***
	(0.464)	(0.510)	(0.371)	(0.359)	(0.201)	(0.275)	(0.525)	(0.475)	(0.327)	(0.343)	(0.066)	(0.078)
Number of teachers (one omit.)												
Second teacher	0.859	0.879	0.667	0.378	0.898	1.169	0.429*	0.400	1.315	2.066	0.519	0.533
	(0.342)	(0.507)	(0.345)	(0.226)	(0.243)	(0.335)	(0.180)	(0.230)	(0.803)	(1.827)	(0.210)	(0.248)
Teacher age	0.975	0.965	1.047	1.028	0.970	0.963	0.945	0.891	1.036	0.986	1.015	1.031
	(0.050)	(0.054)	(0.060)	(0.047)	(0.035)	(0.047)	(0.068)	(0.069)	(0.163)	(0.093)	(0.055)	(0.052)
Years teaching	1.027	1.075	0.938	0.957	1.052	1.070	1.032	1.037	0.940	1.001	1.041	1.051
	(0.064)	(0.060)	(0.074)	(0.072)	(0.042)	(0.053)	(0.091)	(0.110)	(0.136)	(0.103)	(0.045)	(0.052)
Years teaching pre-primary in this school	0.985	0.983	1.013	1.011	0.989	0.992	1.077	1.087*	1.030	1.046	0.964	0.978
	(0.021)	(0.028)	(0.019)	(0.027)	(0.023)	(0.027)	(0.047)	(0.046)	(0.055)	(0.034)	(0.048)	(0.043)

	Materials	Materials	Teacher supports	Teacher supports	Pedagogy	Pedagogy	Life skills	Life skills	Foundational skills	Foundational skills	Multi-disciplinary	Multi-disciplinary
Teacher status (teacher omit.)												
Senior teacher	0.619	0.392*	1.048	1.515	1.230	1.122	1.031	1.162	0.995	0.850	0.509	0.566
	(0.188)	(0.170)	(0.937)	(1.368)	(0.362)	(0.338)	(0.763)	(1.032)	(0.630)	(0.503)	(0.208)	(0.178)
Expert teacher	1.678	0.828	1.049	1.682	0.451	0.281	0.941	1.148	0.713	0.669	0.509	0.370
	(1.058)	(0.636)	(0.946)	(1.145)	(0.264)	(0.189)	(0.935)	(1.387)	(0.614)	(0.718)	(0.306)	(0.201)
Differences in date	1.011	0.975	0.948**	1.081	0.783*	0.699*	1.022	1.026	1.212	1.588	1.043	1.033
	(0.009)	(0.062)	(0.019)	(0.068)	(0.088)	(0.104)	(0.058)	(0.041)	(0.392)	(0.673)	(0.087)	(0.077)
Mean asset index		1.194		1.786		0.464		1.628		0.494		0.661
		(0.517)		(1.151)		(0.193)		(0.894)		(0.384)		(0.234)
% Mother secondary		0.991		0.976		1.004		0.993		0.966***		0.989
		(0.016)		(0.016)		(0.006)		(0.010)		(0.009)		(0.009)
% Father secondary		1.022		1.051*		1.008		0.989		1.030		0.989
		(0.025)		(0.021)		(0.013)		(0.015)		(0.026)		(0.012)
% Mother higher ed.		0.990		0.969		0.991		0.996		0.983		1.001
		(0.018)		(0.017)		(0.008)		(0.015)		(0.011)		(0.011)
% Father higher ed.		1.023		1.057*		1.013		1.002		1.018		0.994
		(0.025)		(0.025)		(0.012)		(0.018)		(0.025)		(0.010)
% Mother work		1.001		1.002		0.992		1.002		0.991		1.001
		(0.016)		(0.018)		(0.009)		(0.011)		(0.016)		(0.009)
% Father professional		0.982		0.985*		1.002		1.000		0.997		0.988
		(0.009)		(0.007)		(0.005)		(0.010)		(0.012)		(0.009)
% Father sales		0.986		1.006		1.007		1.027		0.989		0.990
		(0.020)		(0.021)		(0.007)		(0.030)		(0.017)		(0.010)
N (obs.)	331	287	331	258	331	287	310	273	310	273	310	273
Pseudo R-sq.	0.077	0.155	0.174	0.214	0.071	0.126	0.078	0.113	0.062	0.129	0.112	0.133

Source: Authors' calculations

Notes: *p<0.05; **p<0.01; ***p<0.001. Odds ratios in cells, standard errors in parentheses clustered on the enumerator level. Private is a perfect predictor of consistency in the teacher supports model with SES.

5.3 Simulations of quality enhancement action targeting

What would these findings of variable consistency mean in terms of targeting quality enhancement policies? Table 3 presents simulations of targeting different quality enhancement policies undertaken at the class, school, or district level. These simulations estimate the proportion of classes, schools, or districts identified as targets for quality enhancement action in the QAS tool that are also identified as targets for the same policy using the diagnostic research tool.

The first two policies relate to the overall quality level, as identified by the QAS. We use a continuous measure of the average levels to rank schools. The top 10% or 25%[17] of schools might receive recognition, awards, or resources in recognition of their success. The bottom 10% or 25% of schools might be targeted for remedial actions, up to and including closure. At the class level, only 43.5% of classes identified as in the bottom 10% of classes according to the QAS tool are also identified in the bottom 10% of classes according to the diagnostic research tool. In terms of the top 10% of classes, 30.6% of those identified in the top 10% with the QAS tool are identified as in the top 10% with the diagnostic research tool.

Consistency is slightly higher when using a coarser measure. For the top 25% of classes overall identified by the QAS tool, 51.6% were also in the top 25% in the diagnostic research tool. This statistic is 47.4% consistency for the bottom 25%. Overall, even with a coarser measure, substantially different classes would be targeted based on these measures. For overall performance, consistency is sometimes better at the school or district level for coarser measures. Consistency is worse, 25% for the school and district level for the top 10% overall, and worse (40.9%) for the bottom 10% of schools but better for the bottom 10% of districts (83.3%). Consistency for the bottom and top 25% ranges from 83.3% to 100.0% on the district and school level.

The next policies we consider are targeted measures to enhance pedagogy. We look at the classes that are at the bottom (below minimum) level in terms of pedagogy, which might be targeted for training or coaching assistance, or whose teachers might be targeted for termination. Among classes identified as the bottom level of pedagogy on the QAS tool, 53.0% were also so classified by the diagnostic research tool, along with 41.7% on the school and 16.7% on the district level. The reductions on higher levels are likely due to the increasing potential for an additional class to move a school or district out of the very lowest level.

The QAS could act as a needs assessment for infrastructure improvements. We therefore simulate a policy around targeting the bottom (worst) 10% or 25% of classes in terms of safety hazards (the 10% or 25% of classes with the most hazards). Here we see 58.8% of those identified as the 10% most dangerous classes by the QAS tool are also so identified by the diagnostic research tool, 25.0% on the school level, and 33.3% on the district level. For the bottom 25%, the coarser measure again has more consistency (58.8%) although it too has lower consistency on the school (71.1%) and district (66.7%) level.

[17] Because schools may end up with the same QAS scores, while we use a cutoff of the 10th or 25th percentile, the number of schools at or past cutoff may be more than 10% or 25%.

A substantial lack of consistency occurs for which classes have no materials, to potentially target for additional materials or grants. On the class level, 20.5% of those with no materials per the QAS tool are also so identified in the diagnostic research tool, and 33.3% on the school level. There are no school districts with no materials.

One quality enhancement action that would have relatively more consistency would be targeting remediation to weaker students, i.e., following up with the bottom 10% or 25% of classes in terms of letter identification, potentially with reading tutors, literacy coaching, or other literacy enhancement actions. For the bottom 10% in terms of letter identification, there is 61.1% consistency on the class level, 72.7% on the school level, and 66.7% on the district level. For the bottom 25% in terms of letter identification, 79.5% of the classes and 81.5% schools targeted by the QAS tool would also be targeted by the diagnostic research tool, and 100.0% of the same districts.

Table 3. Simulations of quality enhancement policy targeting (percentage of QAS classes, schools, or districts identified as targets for quality enhancement action that are also identified in the diagnostic research tool)

	Top 10% overall	Top 25% overall	Bottom 10% overall	Bottom 25% overall	Bottom level of pedagogy	Bottom 10% hazards	Bottom 25% hazards	No materials	Bottom 10% letters	Bottom 25% letters
Class	0.306	0.516	0.435	0.474	0.530	0.588	0.779	0.205	0.611	0.795
School	0.250	0.988	0.409	0.966	0.417	0.250	0.711	0.333	0.727	0.815
District	0.250	1.000	0.833	0.833	0.167	0.333	0.667	.	0.667	1.000

Source: Authors’ calculations

6 Discussion and conclusions

LMICs have struggled with the quality of education systems, with a learning crisis: learning basic skills is declining even as access has expanded (World Bank 2018a; Le Nestour, Moscoviz, and Sandefur 2022; UNICEF 2022). Regularly assessing learning and acting on this evidence are cornerstones of improving quality (World Bank 2018a). QAS are a critical mechanism to achieve these goals and ultimately to improve quality, particularly at the pre-primary level, when there are no national assessments to draw on (World Bank 2013; Raikes, Neuman, and Burton 2019). There has, however, been limited evidence on how QAS work in practice in LMICs. This paper investigated how consistently QAS measure quality, and lessons learned on how QAS should be used in LMICs, in light of their measurement challenges.

6.1 Summary

Leveraging parallel data collection efforts in schools, we investigated the consistency of QAS tool data with more detailed diagnostic research tool data. Consistency varied substantially across items, with almost perfect consistency only for two out of 35 items and substantial consistency for only five items. Measures of foundational skills were some of the items with substantial consistency, suggesting these aspects of child development may perform relatively better as aspects of a QAS. Ten items had moderate consistency, including some infrastructure items, teacher's supports and children's multidisciplinary skills. Other more subjective aspects of the classroom environment, materials being present and used, and almost all the measures of pedagogy had lower consistency. On the class and school levels, summary measures, corresponding to levels of the QAS for different dimensions and overall, were correspondingly inconsistent.

Consistency on QAS dimensions, modeled at the classroom level, showed relatively weak relationships with community, school, class, and teacher characteristics. While it is still possible that measures are *consistently* biased, these characteristics at least do not contribute to additional variability. There were also, importantly, consistent results whether an independent data collection firm staff member or MoETE supervisor was acting as the enumerator. This finding suggests that QAS data collection efforts can be embedded within ministries of education and achieve comparable reliability and validity to external evaluations. Such embedding has important implications for cost, feasibility, and sustainability of QAS systems in LMICs.

Simulations of targeting potential quality enhancement actions using the two different data show relatively low consistency as well. For instance, only 31% of classrooms identified as in the top 10% of classes overall in the QAS tool were also so identified in the diagnostic research tool. Results were somewhat more consistent with coarser targeting (targeting 25% rather than 10%, for instance). Results were not substantially more consistent on the school or district level; in some cases they were more consistent and in other cases less consistent. Identifying the bottom 10% or 25% of students in terms of letter recognition was the most consistent in the simulations, which follows from the high consistency of individual foundational skills items.

Taken together, these results suggest that QAS tools are often subject to substantial measurement error. Even on identical items, inconsistency was often substantial. Some of this inconsistency may be coming from different timing of data collection. However, timing will vary at scale as well since different schools would be observed on different days. Different choices in the design and implementation of tools for QAS may also yield fundamentally different decisions on where and how to intervene. For instance, different scales may have contributed to some differences in items that were not identical.

These findings emphasize the importance of QAS tools being carefully designed and tested before implementation, and further, that tools should be calibrated based on their ability to capture information with policy relevance. The results also underscore that QAS tools should not be designed or implemented for high-stakes, summative decisions about specific schools or teachers. Instead, QAS tools can, depending on feasibility and funding, inform national action or act as a formative starting point for quality enhancement. For example, one approach is to use a QAS tool to identify schools that should be followed up with additional observation, rather than making decisions based on QAS tool results alone. We discuss these options in greater detail below.

6.2 Limitations

There are several important limitations to be mindful of when considering the implications of our results. We are examining only one country, one set of standards, and two tools. Results could differ substantially with different countries' standards and tools. In particular, the types of items included in standards and tools appear to have an important relationship with consistency, so variation in the emphasis of standards, e.g., whether they focus on child development outcomes, a quality environment, child-teacher interactions, or other aspects of quality will shape consistency.

Our measures of consistency also embed a number of different issues. We cannot separate test-re-test differences[18] from inter-rater reliability or that questions were, in some cases, asked or scaled differently across the tools. Enumerators were also only sometimes observing the same period, although visiting around the same time. Implementing QAS at scale, this problem is likely to be worse, as children and classes at different schools could be observed at time points that are even further apart. We also do not and cannot know whether the QAS tool or diagnostic tool data, or neither, are ultimately "right or wrong." Although the underlying tools, such as the MELQO tools, have undergone extensive design, testing, and adaptation (Raikes et al. 2019), measuring ECD and quality are fundamentally difficult tasks (Burchinal 2018).

6.3 Policy implications

Our findings on the limited consistency of QAS tool results have important implications for the design of QAS, and especially subsequent quality enhancement actions. One potential approach to QAS, which may be a good starting point for fiscally constrained countries working to initially develop a QAS system, would be to start the QAS as a national-level system, using a random sample of schools to identify quality problems, nationally. The diagnostic research tool and QAS tool data were most consistent at the national level in identifying similar strengths and weaknesses in KG quality in Egypt. Policy makers often do not know the specific quality issues their schools face, or the depths of the learning crisis, and so a nationally representative sample can be a good starting point for a QAS (Wiseman 2014). National quality enhancement actions, such as improvements to the curriculum or nation-wide training programs (e.g., teacher education and induction training) can then be revised to target key quality issues and enhance quality.

Where we did find higher levels of consistency was in the similar national results across tools, as there were similar national means on the QAS tool and diagnostic research tool. An important implication of this finding is that shorter, simpler monitoring tools (which are more feasible and affordable for governments to implement) are as good as more detailed research tools for measuring quality, on average, nationally. This has implications for sustainability in both bringing such data collection within the purview of governments (rather than research institutes or other organizations) and in financing data collection on quality. Since the training was one-

[18] Another potential issue would be whether enumerators successfully observed and entered data for the exact same teacher, classroom, or child. Although there were detailed training instructions on identifying the class list, children, etc., together, there may have been deviations in the field. However, if these were major, we would expect much more consistency on the school level, which is not what we observe.

third the length, and one enumerator did the work of a team of three with the monitoring tools, the QAS tools were roughly a third of the staff costs to implement as the diagnostic research tools (see appendix 2). Given the relative consistency between data collection firm and MoETE supervisors, who were randomly assigned, one important implication of our results is that existing staff may well be able to undertake QAS data collection; although there may be some opportunity cost in shifting their time allocation to the QAS, this may make implementation and eventual scale-up more feasible.

While ministry staff can be used for cost savings, other aspects of the QAS system may merit more investment. Longer observation times, for example, may lead to more consistent results. Ongoing efforts to validate and improve question and response design may make the QAS more informative. Increasing the length of training, as well as the quality of training are important areas that may improve the quality of the QAS data itself. Which of these improve consistency and accuracy most cost-effectively is an important question that merits further research. One challenge in Egypt was that we were unable to obtain local videos of teaching in classrooms, due to security and privacy concerns, so enumerators trained on international examples. Investing in local training materials may help improve the functioning of the QAS. The enumerators all did pass reliability quizzes (80% correct or above), but this may not be sufficient to ensure high reliability during fieldwork.

Collecting a full census of QAS tool data from all schools on a regular basis can also be valuable, but may be costly, and resulting data should be used with caution when informing quality enhancement actions. QAS tools should be low-stakes, not high-stakes, and treated as formative not summative. There can sometimes be positive potential information/incentive effects of quality information being released, as with school report cards (Aturupane et al. 2014; Andrabi, Das, and Khwaja 2017; Bassok, Dee, and Latham 2019). We recommend against using QAS results for sanctions or rewards on a teacher level (raises, promotion, bonuses, retention, or firing), given their imprecision. Likewise, we recommend against rewards or sanctions on a school level for performance. Recognition (e.g., certificates, banners) for top performing teachers or schools may be inaccurate but also may incentivize quality (Cotofan 2021).

When governments lack the resources to provide interventions universally, the QAS can help target interventions. Although imprecise, the information in the monitoring tool is nonetheless an upgrade from not knowing how to target at all. For some interventions, such as hiring additional teachers to reduce class size, other systems (e.g., enrollment and staffing data) may allow equally good or better targeting. For other interventions, such as reading tutors or literacy training, knowing which schools have more children struggling with literacy can be helpful for targeting. QAS results may be too imprecise to accurately target professional development or coaching for teachers. Yet, they could act as a starting point for activities such as coaching (which then includes subsequent observations as well, providing further data on quality). However, implementing tailored coaching at scale can also be quite difficult.

In terms of QAS design, while children's skills are naturally variable, they do seem to be one of the better-measured items in QAS. Especially since links between observed quality and child outcomes have been found to be weak in QRIS (Zellman et al. 2008; Elicker et al. 2011; Tout et al. 2011; Keys et al. 2013), directly including standards and monitoring children's skills is also an important part of QAS design. It is, however, important to keep in mind that these skill outcomes are a function of not just the school, and so should inform support and quality enhancement activities, not sanctions or rewards. Ultimately, if QAS are successful, they will incentivize schools to meet quality standards and provide quality enhancement, nationally or in a targeted fashion. A key implication is thus ensuring that standards and the monitoring tool accurately reflect the central goals of the education system, which may vary by school level or grade.

7 References

Anderson, Kate, Abbie Raikes, Sunita Kosaraju, and Alex Solano. 2017. National Early Childhood Care and Education Quality Monitoring Systems. Center for Universal Education at Brookings.

Andrabi, Tahir, Jishnu Das, and Asim Ijaz Khwaja. 2017. Report Cards: The Impact of Providing School and Child Test Scores on Educational Markets. *American Economic Review* 107 (6): 1535–1563.

Araujo, M. Caridad, Pedro Carneiro, Yyannu Cruz-Aguayo, and Norbert Schady. 2016. Teacher Quality and Learning Outcomes in Kindergarten. *Quarterly Journal of Economics* 131 (3): 1415–1454.

Aturupane, Harsha, Paul Glewwe, Renato Ravina, Upul Sonnadara, and Suzanne Wisniewski. 2014. An Assessment of the Impacts of Sri Lanka's Programme for School Improvement and School Report Card Programme on Students' Academic Progress. *Journal of Development Studies* 50 (12): 1647–1669.

Bassok, Daphna, Thomas Dee, and Scott Latham. 2019. The Effects of Accountability Incentives in Early Childhood Education. *Journal of Policy Analysis and Management* 38 (4): 836–866.

Bendini, Magdalena, and Amanda E. Devercelli, ed. 2022. *Quality Early Learning: Nurturing Children's Potential*. Washington, D.C.: World Bank.

Berlinski, Samuel, Sebastian Galiani, and Paul Gertler. 2009. The Effect of Pre-Primary Education on Primary School Performance. *Journal of Public Economics* 93: 219–234.

Berlinski, Samuel, Sebastian Galiani, and Marco Manacorda. 2008. Giving Children a Better Start: Preschool Attendance and School-Age Profiles. *Journal of Public Economics* 92: 1416–1440.

Blimpo, Moussa P., Pedro Carneiro, Pamela Jervis, and Todd Pugatch. 2019. Improving Access and Quality in Early Childhood Development Programs: Experimental Evidence from The Gambia. *GLO Discussion Paper No.* 318. Maastricht, The Netherlands.

Blimpo, Moussa P., and David K Evans. 2011. School-Based Management and Educational Outcomes: Lessons from a Randomized Field Experiment. *Unpublished Manuscript*.

Bouguen, Adrien, Deon Filmer, Karen Macours, and Sophie Naudeau. 2018. Preschool and Parental Response in a Second Best World: Evidence from a School Construction Experiment. *Journal of Human Resources* 53 (2): 474–512.

Bound, John, Charles Brown, and Nancy Mathiowetz. 2001. Measurement Error in Survey Data. *Handbook of Econometrics* 5: 3705–3843.

Brunette, Waylon, Samuel Sudar, Mitchell Sundt, Clarice Larson, Jeffrey Beorse, and Richard Anderson. 2017. Open Data Kit 2.0: A Services-Based Application Framework for Disconnected Data Management. *MobiSys 2017 - Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*.

Burchinal, Margaret. 2018. Measuring Early Care and Education Quality. *Child Development Perspectives* 12 (1): 3–9.

Conn, Katharine M. 2017. Identifying Effective Education Interventions in Sub-Saharan Africa: A Meta-Analysis of Rigorous Impact Evaluations. *Review of Educational Research* 20 (10): 1–36.

Cotofan, Maria. 2021. Learning from Praise: Evidence from a Field Experiment with Teachers. *Journal of Public Economics* 204: 104540.

Cristia, Julian P., Pablo Ibarraran, Santiago Cueto, Ana Santiago, and Eugenio Severin. 2012. Technology and Child Development: Evidence from the One Laptop per Child Program. *IDB Working Paper Series No.* 304.

Desimone, Laura, Michael S. Garet, Beatrice F. Birman, Andrew Porter, and Kwang Suk Yoon. 2003. Improving Teachers' in-Service Professional Development in Mathematics and Science: The Role of Postsecondary Institutions. *Educational Policy* 17 (5): 613–649.

ECD Measure. 2022. Brief Early Childhood Quality Measure (BEQI). Retrieved May 16, 2022. https://www.ecdmeasure.org/beqi/.

El-Kogali, Safaa El Tayeb, and Caroline Krafft. 2015. *Expanding Opportunities for the Next Generation: Early Childhood Development in the Middle East and North Africa*. Washington, DC: World Bank.

Elicker, James, Carolyn Clawson Langill, Karen Ruprecht, Joellen Lewsader, and Treshawn Anderson. 2011. Evaluation of "Paths to QUALITY," Indiana's Child Care Quality Rating and Improvement System: Final Report (Technical Report #3).

Evans, David, and Amina Mendez Acosta. 2021. Education in Africa: What Are We Learning?. *Journal of African Economies* 30 (1): 13–54.

Evans, David, and Anna Popova. 2016. What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews. *The World Bank Research Observer* 31 (2): 242–270.

Fernald, Lia C. H., Elizabeth Prado, Patricia Kariger, and Abbie Raikes. 2017. A Toolkit for Measuring Early Childhood Development in Low- and Middle-Income Countries. Washington, D.C.: World Bank.

Fleisch, Brahm, Volker Schöer, Gareth Roberts, and Amy Thornton. 2016. System-Wide Improvement of Early-Grade Mathematics: New Evidence from the Gauteng Primary Language and Mathematics Strategy. *International Journal of Educational Development* 49: 157–174.

Ganimian, Alejandro J., and Richard J. Murnane. 2016. Improving Educational Outcomes in Developing Countries: Lessons from Rigorous Impact Evaluations. *Review of Educational Research* 86 (3): 719–755.

Glewwe, Paul, Eric A. Hanushek, Sarah Humpage, and Renato Ravina. 2013. School Resources and Educational Outcomes in Developing Countries: A Review of the Literature from 1990 to 2010. In *Education Policy in Developing Countries*, edited by Paul Glewwe, 13–64. Chicago, IL: University of Chicago Press.

Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2010. Teacher Incentives. *American Economic Journal: Applied Economics* 2 (3): 205–227.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin. 2009. Textbooks and Test Scores: Evidence from a Randomized Evaluation in Kenya. *American Economic Journal: Applied Economics* 1 (1): 112–135.

Glewwe, Paul, Michael Kremer, Sylvie Moulin, and Eric Zitzewitz. 2004. Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya. *Journal of Development Economics* 74 (1): 251–268.

Glewwe, Paul, and Eugenie Maïga. 2011. The Impacts of School Management Reforms in Madagascar: Do the Impacts Vary by Teacher Type ?. *Journal of Development Effectiveness* 3 (4): 435–469.

Glewwe, Paul, and Karthik Muralidharan. 2016. Improving School Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications. In *Handbook* *of the Economics of Education*, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, 5:653–743. Amsterdam.

Goodman, Sarena F., and Lesley J. Turner. 2013. The Design of Teacher Incentive Pay and Educational Outcomes: Evidence from the New York City Bonus Program. *Journal of Labor Economics* 31 (2): 409–420.

Gwet, Kilem L. 2014. *Handbook of Inter-Rater Reliability*. Fifth Edit. Vol. 1. AgreeStat Analytics.

Keys, Tran D., George Farkas, Margaret R. Burchinal, Greg J. Duncan, Deborah L. Vandell, Weilin Li, Erik A. Ruzek, and Carollee Howes. 2013. Preschool Center Quality and School Readiness: Quality Effects and Variation by Demographic and Child Characteristics. *Child Development* 84 (4): 1171–1190.

Kotze, Janeli, Brahm Fleisch, and Stephen Taylor. 2019. Alternative Forms of Early Grade Instructional Coaching: Emerging Evidence from Field Experiments in South Africa. *International Journal of Educational Development* 66: 203–213.

Krafft, C., A. Elbadawy, and M. Sieverding. 2019. Constrained School Choice in Egypt. *International Journal of Educational Development* 71 (102104).

Krafft, Caroline, Abbie Raikes, Samira Nikaein Towfighian, and Rebecca Sayre Mojgani. 2023. Quality and Inequality in Pre-Primary and Home Environment Inputs to Early Childhood Development in Egypt. *World Bank Policy Research Working Paper Series No.* 10317. Washington, D.C.

Krishnaratne, Shari, Howard White, and Ella Carpenter. 2013. Quality Education for All Children? What Works in Education in Developing Countries. *3ie Working Paper No.* 20.

Landis, J. Richard, and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33: 159–174.

Le Nestour, Alexis, Laura Moscoviz, and Justin Sandefur. 2022. The Long-Run Decline of Education Quality in the Developing World. *Center for Global Development Working Paper Series No.* 608.

McCoy, Dana Charles, Marcus Waldman, CREDI Field Team, and Günther Fink. 2018. Measuring Early Childhood Development at a Global Scale: Evidence from the Caregiver-Reported Early Development Instruments. *Early Childhood Research Quarterly* 45: 58–68.

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

Ministry of Planning and Economic Development. 2015. Egypt Vision 2030.

———. 2018. Egypt′s Voluntary National Review 2018.

Molina, Ezequiel, Syeda Farwa Fatima, Andrew Ho, Carolina Melo Hurtado, Tracy Wilichowksi, and Adelle Pushparatnam. 2018. Measuring Teaching Practices at Scale: Results from the Development and Validation of the Teach Classroom Observation Tool. *World Bank Policy Research Working Paper Series No.* 8653.

Moustafa, Nariman, Ebtehal Elghamrawy, Katherine King, and Yu (Claire) Hao. 2022. Education 2.0: A Vision for Educational Transformation in Egypt. In *Education to Build Back Better*, edited by Fernando M. Reimers, Uche Amaechi, Alysha Banerji, and Margaret Wang, 51–74. Cham, Switzerland: Springer.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2011. Teacher Performance Pay: Experimental Evidence from India. *Journal of Political Economy* 119 (1): 39–77.

Neal, Derek. 2011. The Design of Performance Pay in Education. *NBER Working Paper Series*

No. 16710. Cambridge, MA.

Pallante, Daniel H., and Young Suk Kim. 2013. The Effect of a Multicomponent Literacy Instruction Model on Literacy Growth for Kindergartners and First-Grade Students in Chile. *International Journal of Psychology* 48 (5): 747–761.

Piper, Benjamin, Stephanie Simmons Zuilkowski, Margaret Dubeck, Evelyn Jepkemei, and Simon J. King. 2018. Identifying the Essential Ingredients to Literacy and Numeracy Improvement: Teacher Professional Development and Coaching, Student Textbooks, and Structured Teachers' Guides. *World Development* 106: 324–336.

Popova, Anna, David K. Evans, Mary E. Breeding, and Violeta Arancibia. 2022. Teacher Professional Development around the World: The Gap between Evidence and Practice. *World Bank Research Observer* 37 (1): 107–136.

Pradhan, Menno, Daniel Suryadarma, Amanda Beatty, Maisy Wong, Arya Gaduh, Armida Alisjahbana, and Rima Prama Artha. 2014. Improving Educational Quality through Enhancing Community Participation: Results from a Randomized Field Experiment in Indonesia. *American Economic Journal: Applied Economics* 6 (2): 105–126.

Raikes, Abbie, Natalie Koziol, Dawn Davis, and Anna Burton. 2020. Measuring Quality of Preprimary Education in Sub-Saharan Africa: Evaluation of the Measuring Early Learning Environments Scale. *Early Childhood Research Quarterly* 53 (4): 571–585.

Raikes, Abbie, Michelle Neuman, and Anna Burton. 2019. Quality Standards and Quality Assurance Systems for Pre-Primary Education. *UNICEF White Paper*. New York, NY.

Raikes, Abbie, Rebecca Sayre Mojgani, Jem Heinzel-Nelson Alvarenga Lima, Dawn Davis, Cecelia Cassell, Marcus Waldman, and Elsa Escalante. 2023. Profiles of Quality in Three Distinct Early Childhood Programs Using the Brief Early Childhood Quality Inventory (BEQI). *International Journal of Early Childhood*.

Raikes, Abbie, Rebecca Sayre, Dawn Davis, Kate Anderson, Marilou Hyson, Evelyn Seminario, and Anna Burton. 2019. The Measuring Early Learning Quality & Outcomes Initiative: Purpose, Process and Results. *Early Years* 39 (4): 360–375.

Raikes, Abbie, Rebecca Sayre, and Jem Heinzel-Nelson Alvarenga Lima. 2021. Early Childhood Care & Education Quality Assurance Systems in Africa. USAID and ECD Measure.

Reinke, Wendy M., Melissa Stormont, Keith C. Herman, and Lori Newcomer. 2014. Using Coaching to Support Teacher Implementation of Classroom-Based Interventions. *Journal of Behavioral Education* 23: 150–167.

Rolla San Francisco, Andrea, Melissa Arias, Renata Villers, and Catherine Snow. 2006. Evaluating the Impact of Different Early Literacy Interventions on Low-Income Costa Rican Kindergarteners. *International Journal of Educational Research* 45 (3): 188–201.

Santibanez, Lucrecia, Raul Abreu-Lastra, and Jennifer L. O. Donoghue. 2014. School Based Management Effects: Resources or Governance Change? Evidence from Mexico. *Economics of Education Review* 39: 97–109.

Sayre, Rebecca, Abbie Raikes, and Amanda Devercelli. 2018. Tanzania Pre-Primary Workforce. World Bank.

Tout, Kathryn, Rebecca Starr, Tabitha Isner, Jennifer Cleveland, Ladia Albertson-Junkans, Margaret Soli, and Katie Quinn. 2011. Evaluation of Parent Aware: Minnesota's Quality Rating and Improvement System Pilot. Minnesota Early Learning Foundation.

UNESCO. 2017. Overview of MELQO: Measuring Early Learning Quality Outcomes. Paris: UNESCO, UNICEF, World Bank, & Brookings Institution.

UNICEF. 2022. The State of Global Learning Poverty: 2022 Update.

Westbrook, Jo, Naureen Durrani, Rhona Brown, David Orr, John Pryor, Janet Boddy, and Francesca Salvi. 2013. Pedagogy, Curriculum, Teaching Practices and Teacher Education in Developing Countries. *Final Report. Education Rigorous Literature Review.*

Wiseman, Alexander W. 2014. Policy Responses to PISA in Comparative Perspective. In *PISA, Power, and Policy: The Emergence of Global Educational Governance,* edited by Heinz-Dieter Meyer and Aaron Benavot, 303–322. Oxford, UK: Symposium Books.

Wolf, Sharon. 2018. Impacts of Pre-Service Training and Coaching on Kindergarten Quality and Student Learning Outcomes in Ghana. *Studies in Educational Evaluation* 59: 112–123.

Wolf, Sharon, J. Lawrence Aber, Jere R. Behrman, and Edward Tsinigo. 2019. Experimental Impacts of the "Quality Preschool for Ghana" Interventions on Teacher Professional Well-Being, Classroom Quality, and Children's School Readiness. *Journal of Research on Educational Effectiveness* 12 (1): 10–37.

Wolf, Sharon, Mahjabeen Raza, Sharon Kim, J. Lawrence Aber, Jere Behrman, and Edward Seidman. 2018. Measuring and Predicting Process Quality in Ghanaian Pre-Primary Classrooms Using the Teacher Instructional Practices and Processes System (TIPPS). *Early Childhood Research Quarterly* 45: 18–30.

World Bank. 2023. Education Statistics - Policy Data. *SABER-ECD*. Accessed January 5. https://datatopics.worldbank.org/education/wDashboard/dqpolicy.

———. 2013. What Matters Most for Early Childhood Development: A Framework Paper. *SABER Working Paper Series No.* 5. Washington, DC.

———. 2018a. Learning to Realize Education's Promise. Washington, DC: World Bank.

———. 2018b. Supporting Egypt Education Reform Project. Project Appraisal Document. Washington, D.C.: World Bank.

———. 2022. World Development Indicators. *World Bank Databank*. Retrieved January 6, 2022. https://databank.worldbank.org/source/world-development-indicators#.

Yoon, Kwang Suk, Teresa Duncan, Silvia Wen-Yu Lee, Beth Scarloss, and Kathy L. Shapley. 2007. Reviewing the Evidence on How Teacher Professional Development Affects Student Achievement. *Issues & Answers* (REL 2007 No. 033).

Zellman, Gail L., Michal Perlman, Vi-Nhuan Le, and Claude Messan Setodji. 2008. *Assessing the Validity of the Qualistar Early Learning Quality Rating and Improvement System as a Tool for Improving Child Care Quality*. RAND Corporation.

8 Appendix 1: Standards and levels

Dimension	Sub-dimension	Std. #	Standard	Levels of performance	Indicators
Infrastructure and Materials **Overarching standard: KGs provide space and materials that ensure that children can safely and securely play and learn.**	1.Infrastructure	1.1	Desk/table work space: Children have safe, child-sized work spaces (desks or tables and seats) with sufficient room to do their work	Below minimum: Class size > 50 OR Student-teacher ratio > 36 OR No on any of (1) desk/table work space, (2) toileting facilities, (3) handwashing facilities, (4) drinking water facilities, (5) lighting, (6) ventilation, (7) hazard Minimum**:** Class size <=50 AND Student-teacher ratio <=36 AND Yes on all of (1) desk/table work space, (2) toileting facilities, (3) handwashing facilities, (4) drinking water facilities, (5) lighting, (6) ventilation, (7) hazard Developing**:** + Activity space in the classroom + Activity space outside the classroom Achieved**:** + class size <=36 AND student-teacher ratio <=18	- Desks present and sound - Seating present and sound
		1.2	Activity space in the classroom: Children have sufficient space within the classroom to leave their seats and undertake activities		- Activity space in the classroom sufficient
		1.3	Activity space outside the classroom: Children have activity spaces outside the classroom (e.g. activity room for music, drama, etc., gym, garden, or playground)		- Activity space outside the classroom exists
		1.4	Toileting facilities: Children have access to sex-appropriate, size appropriate, and clean toileting facilities		- Toileting facilities exist and are sex- and size-appropriate - Clean toileting facilities
		1.5	Handwashing facilities: Children have access to size-appropriate, clean handwashing facilities, with running water and soap		- Handwashing facilities exist - Handwashing facilities have running water - Handwashing facilities are size appropriate - Handwashing facilities have soap
		1.6	Drinking water facilities: Children have access to size-appropriate, clean drinking water facilities, with running water		- Drinking water facilities exist - Drinking water facilities have running water - Drinking water facilities are clean
		1.7	Lighting: KG classes have adequate lighting for all children to easily see and read		- Adequate lighting present
		1.8	Ventilation: KG classes have adequate ventilation for all children		- Adequate ventilation present

Dimension	Sub-dimension	Std. #	Standard	Levels of performance	Indicators
		1.9	Free from hazards: Classroom and school are free from hazards to children's safety		- Hazards absent
		1.10	Class size: Each KG class has a maximum of 36 students		- Class size (enrolled)
		1.11	Student/teacher ratio: The maximum student/teacher ratio in a class is 16		-Student/teacher ratio
	2.Materials	2.1	Writing and drawing tools: The classroom has and the children use age-appropriate writing and drawing tools	Below minimum: Missing or not using writing and drawing tools Minimum: Using writing and drawing tools Developing: + classroom management tools Achieved: + manipulatives	- Writing tools present (pencils, crayons, etc.) - Writing tools used by children
		2.2	Manipulatives: The classroom has and the children use manipulatives to facilitate learning		- Manipulatives present (clay, straws, blocks, etc.) - Manipulatives used by children to learn
		2.3	Classroom management: The classroom has, and the children engage with, tools for classroom management to help organize the children's schedule		-Classroom management tools present (calendars, charts, clocks, etc.) -Children engage with classroom management tools (Classroom management tools used to organize children)
Teachers and Pedagogy 2.0 **Overarching standard: Teachers are credentialed, trained, and supported so that they can deliver child-centered education that is relevant to children's lives.**	3.Teacher's supports	3.1	Degree in Early Childhood Studies: The teacher has a degree in Early Childhood Studies	Below minimum: Teacher does not have a degree in early childhood studies Minimum: Teacher has a degree in early childhood studies Developing: + at least one of (1) training on education 2.0 or (2) continuous professional development Achieved: + both (1) training on education 2.0 and (3) other continuous professional development	- Degree in early childhood studies
		3.2	Training on Education 2.0 Curriculum & Pedagogy: The teacher received in-service training on the curriculum and pedagogy 2.0		- Received training on curriculum and pedagogy 2.0
		3.3	Continuous professional development: The teacher has undertaken five days of continuous professional development/education within the past 12 months		- Attended 5 days of professional development
		3.4	Supervisor support: The KG supervisor supports teachers' self-development and professional development in		

Dimension	Sub-dimension	Std. #	Standard	Levels of performance	Indicators
			order to ensure appropriate curriculum and pedagogy 2.0 implementation		
		3.5	School principal support: The school principal supports the KG teacher and class with appropriate resources to ensure all children have a safe and conducive environment to learn.		
	4. Pedagogy 2.0	4.1	Lesson plan: The teacher has and follows the education 2.0 lesson plan	Below minimum: Teaching is not child centered or not open ended Minimum: Teaching is child-centered and is open-ended Developing: + Teaching is relevant to everyday life Achieved: + Teaching uses the strategies of education 2.0	- Teacher uses the strategies of education 2.0
		4.2	Teacher's guide: The teacher has and follows the teacher's guide		- Teacher uses the strategies of education 2.0
		4.3	Child-centered: Children are given appropriate support and feedback to complete tasks. The teacher engages with children with warm and responsive interactions and provides appropriate supervision for guiding behaviors.		- Teacher gives individualized instruction - Teacher facilitates positive interactions - Teacher encourages appropriate behavior and redirects inappropriate behavior
		4.4	Play-based: Teaching is play-based and interactive, offering choice to engage children		- Teacher uses the strategies of education 2.0
		4.5	Project-based: Teaching is project-based, engaging children in projects that reflect the real world and use real materials		- Teacher uses the strategies of education 2.0
		4.6	Open-ended: Dialogue between teachers and students is open-ended, using discussion, back and forth dialogue, and open-ended questions rather than rote memorization.		- Dialogue between teachers and children is open ended - Teacher uses discussion
		4.7	Relevant to everyday life: The teacher connects lessons to		- Teachers relate learning to everyday life

Dimension	Sub-dimension	Std. #	Standard	Levels of performance	Indicators
			everyday life and uses examples from everyday life		
Child learning and development **Overarching standard: Children learn foundational skills (literacy and numeracy) and life skills across multiple disciplines.**	5. Life skills	5.1	Learning Skills: Children begin to develop key learning skills, including creativity, critical thinking, and problem-solving	Below minimum: The majority of children have developed no life skills Minimum: The majority of children have developed at least one life skill Developing: The majority of children have developed at least two life skills Achieved: The majority of children have developed three life skills	- Children demonstrate learning skills and ability to focus on tasks.
		5.2	Personal empowerment: Children begin to develop personal empowerment skills, including self-management, resilience, and communication		- Children can manage complex instructions
		5.3	Active citizenship: Children begin to develop active citizenship in their classrooms and communities, including respect for diversity, empathy, and participation		- Children develop active citizenship and empathy/consideration
	6. Foundational skills	6.1	Reading, listening, and speaking: Children learn foundational reading, listening, and speaking skills, including letter and basic word recognition and key early vocabulary	Below minimum: Fewer than half of children recognize the majority of letters OR fewer than half of children can write their name OR fewer than half of children can count to 10 Minimum: At least half of children recognize the majority of letters AND half of children can write their name AND half of children can count to 10 Developing: At least 75% of children recognize the majority of letters AND 75% of children can write their name AND 75% of children can count to 10 Achieved: All children recognize the majority of letters AND all of children can write their name AND all children can count to 10	- Children recognize letters
		6.2	Writing: Children acquire foundational writing skills, including writing both individual letters and basic words		- Children can write their name
		6.3	Mathematics: Children acquire basic numeracy and mathematics skills, including basic counting, addition, and subtraction		- Children can count to ten
	7. Multidisciplinary	7.1	Who am I?: Children develop a sense of identity and their relationships with family and school community	Below minimum: Fewer than half of children know school name OR fewer than half of children know country name OR fewer	- Children know school name

Dimension	Sub-dimension	Std. #	Standard	Levels of performance	Indicators
		7.2	The world around me: Children develop a sense of their place in the world around them, including their local community and country	than half of children know where fish live OR fewer than half of children know flag Minimum: At least half of children know school name AND half of children know country name AND half of children know where fish live AND at least half of children know flag Developing: At least 75% of children know school name AND 75% of children know country name AND 75% of children know where fish live AND at least 75% of children know flag Achieved: All children know school name AND All children know country name AND all children know where fish live AND all children know flag	- Children know name of county
		7.3	How does the world work? Children develop an early sense of how the world works in both scientific (e.g. plants need light to grow) and occupational terms (e.g. the work of farmers and doctors).		- Children understand habitats of different animals
		7.4	Communication: Children develop a sense of communication through different mediums (art, music) as well as the importance of communication to relationships and friendships		- Children recognize flag of Egypt

Appendix 2: Features of the Diagnostic Research Tool and QAS Tool

	Features	Diagnostic research tool	QAS tool
Design of the tool	Measuring KG Quality	***A total of 117 items administered through 3 instruments, as follows:*** **Teacher Interview: 45 items asked through survey** • Teacher experience, qualifications, compensation: 9 items • Teacher attitude/motivation: 10 items • Professional development experiences: 4 items • Understanding and attitude about Education 2.0: 10 items • Teacher approach to curriculum and language: 6 items • Safety: 2 items • Past school year: 4 items **Director Interview: 22 items asked through survey** • School and pre-primary information: 5 items • Teacher characteristics: 4 items • Water, sanitation, hygiene (including COVID-19): 4 items • Understanding and attitude of Education 2.0: 9 items **Classroom Observation: 50 items** • Classroom information: 10 items • Learning activities: 13 items • Classroom interactions and approaches to learning: 6 items • Classroom arrangement, space and materials: 16 items • Facilities and safety: 5 items	***A total of 26 items administered through 2 instruments, as follows:*** **Teacher interview: 7 items asked through survey** • Teacher and classroom information (number of teachers/students, teacher education and experience, in-service training) **Classroom Observation: 19 items on yes/no scale** • Education & learning activities (materials and pedagogy): 10 items • Facilities and safety: 9 items
	Measuring KG Learning	***A Total of 152 items administered through 3 instruments, as follows:*** **Child Direct Assessment**: 24 items • 9 items language/literacy • 6 items math/numeracy • 5 items executive function/social-emotional • 4 items multidisciplinary **Teacher Report of Child Development:** 46 items asked through survey • Background: 9 items • Child's social-emotional development: 21 items • Child's cognitive development (math/language): 16 items	***A total of 13 items administered through 2 instruments, as follows:*** **Child Direct Assessment: 10 items** (letter identification, name writing, counting, shapes, multi-disciplinary) **Teacher Report of Child Development:** 3 yes/no items • Social-emotional development: 1 item • Executive function skills: 2 items

	Features	Diagnostic research tool	QAS tool
		Parent/Caregiver Report of Child's learning, SES, home environment: 82 items asked through phone interview • Family background and education: 19 items • Home learning environment and ECD: 12 items • Child's social-emotional development: 22 items • Child's cognitive development (math/language): 20 items • Household characteristics: 9 items	
Administration of the tool	Designed duration, for each KG classroom	• 30 minutes with the director (once per school) ***6.75 hours overall per class*** • 15 minutes for teacher interview • 2.5 hours of classroom observation • 60 minutes for teacher to report on child development for 4 children in classroom (15 minutes per child) • 120 minutes for child assessment of 4 children (30 minutes per child) • 60 minutes for interview with 4 parents (15 minutes per parent)	***2.5 hours overall duration per class*** • 5 minutes for teacher interview • 2 hours of classroom observation • 10 minutes for teacher report on learning for 4 children (2-3 minutes per child) • 40 minutes for child assessment of 4 children (10 minutes per child)
	Responsibility for administration	• The tool is designed to be administered by trained enumerators of an independent data collection firm • The tool requires a team of three enumerators (supervisor; direct assessment and teacher report enumerator; classroom observation and teacher interview enumerator) per school visit for its administration	The tool is designed to be administered by government officials (e.g., supervisors) as part of their routine visits to schools The tool requires one supervisor per school visit for its administration
	Required training	10 days of training, including site visits; enumerators required to pass a written quiz regarding tool content and training procedures and a classroom video quiz with at least 80% agreement with master codes	3 days reliability training, including site visits and written quizzes
	Format	Tablet-based	Tablet-based (designed to allow for paper-based administration too)

	Features	Diagnostic research tool	QAS tool
	Estimated costs	Training: $74/day of training per enumerator, 10 days of training per enumerator, 3 enumerators/supervisors per school ($259 per school) Data collection: $944/school	Training: $96/day of training per enumerator, 3 days of training, one enumerator per school ($38 per school) Data collection: $345 per school Note: The QAS tool is designed to be embedded in supervisors' routine visits to schools. As such, even though we have calculated the cost per school, we estimate that there would be no (or only a negligible) additional recurrent financial cost for the government.
Summary characteristics	Types of items	269 items in total, out of which: • 15 were identical items and responses between the diagnostic research tool and the QAS tool • 11 were questions on a 4-point response scale	39 items in total, out of which: • 15 were identical items and responses between the diagnostic research tool and the QAS tool • Six were based on the diagnostic's 4-point response scale and turned into yes/no
	Instruments administered	Six instruments in total: • Director interview • Classroom observation • Child direct assessment • Teacher interview • Teacher report of child's development • Parent interview	Four instruments in total: • Classroom observation • Child direct assessment • Teacher interview • Teacher report of child's development
	Subjects interviewed	• One director per school • Up to 3 KG1 and 3 KG2 classes • All teachers (usually 1-2) per class • 4 children per class • 4 parents per class	• One director per school • Up to 3 KG1 and 3 KG2 classes • All teachers (usually 1-2) per class • 4 children per class
Other Considerations	Drawbacks	• 4-point classroom observation scale is harder to become reliable on • Less affordable and feasible for government to implement	• Less detailed • No recourse for government official if not reliable on tool if administration is just part of their job