Early Childhood Research Quarterly 53 (2020) 571–585

Contents lists available at ScienceDirect

Early Childhood Research Quarterly

Measuring quality of preprimary education in sub-Saharan Africa: Evaluation of the Measuring Early Learning Environments scale

Abbie Raikes [a,*], Natalie Koziol [b], Dawn Davis [b], Anna Burton [b]

[a] University of Nebraska Medical Center, United States
[b] University of Nebraska, Lincoln, United States

ARTICLE INFO

Article history:
Received 30 April 2019
Received in revised form 8 April 2020
Accepted 2 June 2020

Keywords:
Africa
Measurement
Preprimary classroom observations
Professional development
Quality in early childhood education

ABSTRACT

Measurement of quality in early childhood education (ECE) helps shape policy and practice, yet few studies have examined the adaptation and resulting psychometric properties of ECE quality measures when used in low- and middle-income countries. This study reports on the adaptation of the Measure of Early Learning Environments scale (MELE-A), developed as part of the Measuring Early Learning Quality & Outcomes (MELQO) Initiative, in one sub-Saharan African country. Beginning with a global "core" of items, MELE-A was adapted to address measurement feasibility and align with cultural context and national standards. The sample included 250 public and private preprimary schools and 979 children from all regions of the country. Three factors were hypothesized to represent empirically documented aspects of quality and were supported by categorical confirmatory factor analysis: health/safety, materials/activities, and teacher/child interaction. Few associations were found between these factors and child development and learning and teacher characteristics; only materials/activities demonstrated significant associations with children's learning, while teacher education was associated with all three factors. Results document the multi-faceted process of adapting tools and the importance of documenting psychometric properties of these adapted tools, to improve accuracy of ECE measurement to inform policy and practice.

Access to early childhood education (ECE) is expanding globally (UNESCO Institute of Statistics, 2019), in service of an overall goal to promote lifelong equity in education, health and well-being by investing in early childhood development. Articulated as part of the United Nations' Sustainable Development Goals (SDGs), Target 4.2 states that by 2030, "all children have access to quality early childhood care, development and preprimary education so that they are ready for primary education." Access to ECE continues to steadily increase (UNESCO Institute of Statistics, 2019), but quality remains a concern in many parts of the world (Britto, Yoshikawa, & Boller, 2011). Access to quality ECE is especially important in low- and middle-income countries (LMIC) where children face the greatest risks to development such as inadequate stimulation, nutrition and health care, endemic disease and poverty (Britto et al., 2017).

Several aspects of young children's environments contribute to quality in ECE, including the physical environment, teacher/child interaction, the content of the curricula, the teachers' skill in scaffolding children's development, and the ability of the teacher to individualize instruction, though measurement tools at present may not fully capture all of these constructs (Britto et al., 2011; Burchinal, 2018). Measurement of ECE is a central element of scaling within countries, and data and measurement play a critical role in achieving the vision of equity articulated by the Sustainable Development Goals (Raikes, Yoshikawa, Britto, & Iruka, 2017). Across all targets, the SDGs outline a measurement agenda; for Target 4.2, focused on ensuring access to quality early childhood development programs, global indicators include child development and access to preprimary education with encouragement to measure other constructs, including quality, at the national and regional levels. However, in many low- and middle-income countries, data on quality in ECE is not consistently available, especially within large samples of typical settings (Raikes, Yoshikawa et al., 2017), which in turn impedes coordinated action to address quality. Measurement forms the basis of monitoring systems to inform governments and stakeholders on the status of their investments in preprimary education; to inform parents; and to support teacher knowledge and education (OECD, 2015; Thornburg et al., 2011). Results from quality measurement in diverse contexts can and should be used to expand developmental science and theories on aspects of learning environments that are critical for

* Corresponding author.
E-mail address: abbie.raikes@unmc.edu (A. Raikes).

https://doi.org/10.1016/j.ecresq.2020.06.001
0885-2006/

children's development (Raikes, Davis, & Burton, 2019; UNESCO, UNICEF, Brookings Institution, & The World Bank, 2017). However, achieving these goals requires contextually relevant and feasible measurement (Britto et al., 2011; Yoshikawa, Wuermli, Raikes, Kim, & Kabay, 2018).

In response to measurement needs as part of Target 4.2 and to support national implementation of ECE, the Measuring Early Learning Quality & Outcomes (MELQO) Initiative was begun by four organizations, UNICEF, UNESCO, World Bank and UNICEF (UNESCO et al., 2017). Beginning with a conceptual framework generated through existing measures and empirical evidence, MELQO's Consortium, comprised of researchers and stakeholders in early childhood development from many countries, created two sets of open-source tools focused on preprimary education addressing child development and learning, and quality in preprimary settings. These tools were designed to facilitate measurement in low- and middle-income countries by integrating commonly-articulated global concepts of ECE quality and child development with local adaptation processes, with emphasis on creating feasible, cost-effective measurement (see UNESCO et al., 2017, for a complete description, and see (Raikes, Sayre, Davis, Anderson, Hyson, Seminario, & Burton, 2019), for information on psychometric evidence supporting scores from other MELQO instruments). This study reports on the development and testing of a recent open-source ECE observational quality measure, the Measure of Early Learning Environments (MELE), as it was adapted and tested in one country in sub-Saharan Africa.

1. Early childhood education quality in low- and middle-income countries

Children's learning and development and country GDP are associated, due to the unique developmental challenges facing children living in low-income countries (Bornstein et al., 2012). Investing in ECE offers considerable promise for addressing the chronic failures of primary education in many low-income countries, where large percentages of children fail to complete primary education (Crouch & Merseth, 2017). ECE has been shown to have positive impacts on children's development and learning across countries (Jackson, Ahmed, Carslake, & Lietz, 2019) even when children only have access to basic ECE provision (Rao et al., 2012). Positive impacts of ECE have been reported in some of the poorest areas of the world, including in East Africa (Bietenbeck, Ericsson, & Wamalwa, 2019; Martinez, Naudeau, & Pereira, 2012); and South Asia (e.g., Aboud & Hossain, 2011; Rao et al., 2012; Singh & Mukherjee, 2018).

Analyses of the associations between ECE and child development across observational and experimental designs report a range of impact estimates, from small (e.g., Bietenbeck et al., 2019; Jackson et al., 2019) to large. For example, Rao, Sun, Chen, and Ip (2017) reported an average effect of 0.70, with larger results for higher-quality interventions, as defined by teacher qualifications, the existence of a child-centered curriculum, and teacher–child ratios. When positive effects of ECE on child learning are not found, the low quality of ECE is mentioned as a potential cause (e.g., Gong, Xu, & Han, 2016), and ECE has been shown to have greater positive impacts as quality increases (e.g, in India, Kaul, Chaudhary, & Sharma, 2014; East Africa, Malmberg, Mwaura, & Sylva, 2011; Ghana, McCoy & Wolf, 2018).

However, there is little research directly documenting quality of ECE in many parts of the world, especially in sub-Saharan Africa, an important focal regional area for early childhood development where ECE could make a substantial contribution (Garcia et al., 2008). Although the African Union, among other leadership entities, has acknowledged the critical role of investing in early childhood development to address lifelong health, learning and well-being, ECE investments are small and inconsistent (Neuman & Okeng'o, 2019), with heavy reliance on the private sector (UNESCO Institute of Statistics, 2019), leading to considerable variation in the types, settings, and quality of ECE. For example, while only half of preschools in Lagos, Nigeria had playgrounds, nearly all preschools in Accra, Ghana and Johannesburg, SA had playgrounds; and while no teachers in Lagos had completed some training in early childhood development, 40% of teachers in Ghana and 85% of teachers in Johannesburg had completed ECD training (Bidwell & Watine, 2014). Similarly, classroom practices vary considerably across countries, with 87% of classroom time spent with young children sitting in rows of desks facing forward in a sample of preschools in Nairobi, while only 18% of time was spent this way in Johannesburg (Bidwell & Watine, 2014). While some studies find that ECE health and safety conditions are adequate, other studies report notable deficiencies in water, sanitation and safety (Kotzé, 2015). Yet some similarities in ECE quality across countries have also emerged; for example, most teachers in Ethiopia rely on rote memorization (Rossiter, Hagos, Rose, Teferra, & Woldehanna, 2018) and in Tanzania, children also had few opportunities for engagement, especially in rural areas (RTI, 2018). Variations in ECE quality may also reflect investments; resources for ECE are limited in sub-Saharan Africa, with just 2% of education budgets devoted to ECE (Jamarillo & Mingat, 2008), although recent investments may have increased this percentage.

2. Evaluating psychometric properties of ECE quality measures

Several sources of validity and reliability evidence can be used evaluate functioning of quality measures (cf., *Standards for Educational and Psychological Testing*, American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME), 2014): (1) Evidence based on test content, or the extent to which test content adequately samples the underlying constructs; (2) Evidence based on internal structure, or the extent to which the data support the hypothesized factor structures; (3) Evidence of internal consistency, or the extent to which responses among items are highly correlated; and (4) Evidence based on relations to other variables, or the extent to which the MELE scores are associated with child outcomes and teacher and school characteristics.

A substantial body of research addresses psychometric properties of ECE quality measures in the US and other high-income countries. ECE quality measurement has become intertwined with ECE policy in the United States (e.g. Early, Sideris, Neitzel, LaForett, & Nehler, 2018) but this heavy reliance on quality measures coexists with concerns about the content, structure and psychometric functioning of measures, including how well measures differentiate across levels of quality and the need for measures that index content of curricula and differentiated instruction, among other limitations (Burchinal, 2018). A review of the evidence of two commonly used scales for measuring ECE classroom quality, the Environmental Ratings Scales (e.g., ECERS; Harms, Clifford, & Cryer, 1980; ECERS-R; Harms, Clifford, & Cryer, 1998; ECERS-E, Sylva, Siraj-Blatchford, & Taggart, 2003; ECERS-3; Harms, Clifford, & Cryer, 2015) and the Classroom Assessment Scoring System (CLASS-PreK; Pianta, La Paro, & Hamre, 2008) are described below to illustrate these issues. Versions of the ECERS, referred to collectively as the Environmental Rating Scales or ERS, cover a wide range of constructs, spanning the physical characteristics of the classroom, family and community engagement, children's interactions with teachers and peers, access to materials, and program structure, while the CLASS is focused specifically on emotional support, materials/activities, and instructional support.

2.1. Test content

In this study, we define test content as the ideas, beliefs and behaviors that comprise definitions of quality in ECE, including those articulated in government policy and those that are commonly held among parents, teachers and others. While locally held quality definitions may encompass ideas beyond government standards and test content can also be defined by theories and science on child development, national government standards outlining content expectations of preprimary education are increasingly common and articulate a vision for ECE quality (Raikes, Davis et al., 2019; OECD, 2015). In the United States, the ECERS and other scales have been directly integrated into early childhood policies, such as Quality Rating and Improvement Systems (e.g., Administration for Children & Families, 2015); policies and practices have become so merged with quality measurement that some researchers have noted that quality is ultimately defined by the ECERS and CLASS (Early et al., 2018). However, test content analyses may be more relevant in other policy contexts, where government policies on quality, including quality standards, may work more interactively with routine and large-scale quality measurement and are less closely merged. As well, test content can also be indexed by its ability to represent the cultural and contextual beliefs that comprise quality.

2.2. Internal structure and internal consistency

Quality measures are intended to capture latent constructs of quality learning environments. As such, items should demonstrate (1) coherence to predicted underlying latent constructs of quality; and (2) internal consistency. While observational ECE quality scores demonstrate evidence of internal consistency as measured by coefficient alpha (α) greater than .70, (e.g., in China, see Hu, Fan, Gu, & Yang, 2016 [CLASS-PreK]; Sandilos & DiPerna, 2014 [CLASS-PreK]; in Sweden and South Korea, see Sheridan, Giota, Han, & Kwon, 2009 [ECERS]), there is less evidence for a consistent factor structure, raising questions regarding the accuracy of these scales. Using a large sample of ECERS-R (Harms et al., 1998) data from the United States, Gordon, Fujimoto, Kaestner, Korenman, and Abner (2013) assessed the specificity of ECERS-R domains for child development. Factor analyses did not identify one underlying factor nor a replication of the six constructs proposed by the test designers, consistent with results from factor analyses from many other studies, and instead found a three-factor solution of items from the space/furnishings/activities/routines subscale; the personal care subscale; and the language reasoning/interactions subscale showed the best model fit.

Similar mixed results have been reported for the ECERS-3 (Harms et al., 2015). Using exploratory analyses, Early et al. (2018) reported four factors (learning opportunities; gross motor; teaching interactions; and math activities). Montes et al. (2018), however, found three clusters of ECERS-3 items (learning activities and environment, interaction, and gross motor related) as well as many items that did not load onto any construct. Inconsistent results have also been reported for the CLASS scale. Some analyses, including a recent meta-analyses, have reproduced the hypothesized three-factor structure for the CLASS albeit with some modifications to the scale (Bihler, Agache, Kohl, Willard, & Leyendecker, 2018; Hamre et al., 2013; Li, Liu, & Hunter, 2019; Pakarinen et al., 2010), but other studies have found better model fit with other solutions, such as a bifactor model that differentiates aspects of teacher/child interactions (Hamre, Hatfield, Pianta, & Jamil, 2014). In sum, factor analyses to date show a high degree of variability in factor structures of quality measures, suggesting that items do not index the underlying constructs accurately and raises the risk of measurement error when relying on subscales to summarize "quality." At the same time, results from various analyses roughly indicate emergence of similar themes addressing aspects of teacher/child interactions; children's exposure to organized classrooms with materials and a curriculum; and the safety of the physical environment.

2.3. Association with other variables

An extensive body of literature addresses validity evidence of CLASS and ECERS in the U.S. and other high-income countries (HIC), especially addressing the strength of associations between quality scores and child development (see Burchinal, 2018, for a review; Burchinal, Kainz, & Cai, 2011; Keys et al., 2013), which is arguably one of the most critical aspects of validity of ECE quality measures. Some research from high-income countries, suggests that the ECERS-R and ECERS-E adequately index quality in ECE settings as defined by expected associations with child development and teacher characteristics (Sylva et al., 2006), while other recent studies, including a meta-analyses, Brunsek et al. (2017) found weak but positive associations between ECERS and ECERS-R total scores and children's language and positive social behaviors, with 14 of 17 tested effects showing non-significance, and an association between language reasoning and language scores. Similar small but significant associations were reported by Early et al. (2018), and Gordon et al. (2013). Recent meta-analyses of the CLASS indicate that classroom organization scores showed small but reliable associations with measures of children's executive functioning (measured by pencil tapping), while instructional support showed small but reliable associations with social skills (Perlman et al., 2016), although 12 of 14 associations tested were not statistically significant.

More highly qualified teachers delivered higher-quality services in several studies (e.g. Burchinal et al., 2008; Early et al., 2007; Hamre, 2014), although with mixed effect sizes. Some studies have shown that teacher education, professional development and experience (usually defined as years of teaching) predict classroom quality as measured by both the ECERS-R and CLASS in U.S. samples (e.g., Denny, Hallam, & Homer, 2012). In contrast, other studies have shown null associations between teacher education and ECE classroom quality (Early et al., 2007), or a negative association between years of teaching and emotional support in the CLASS, which was reported in Finland (Pakarinen et al., 2010). A review of studies from 23 countries (including one low-income country) using versions of the Environmental Rating Scales found that ERS scores, even when various versions and lengths of the instrument were used, were positively associated with child/teacher ratios and caregiver sensitivity (Vermeer, van IJzendoorn, Cárcamo, & Harrison, 2016).

3. Measuring ECE quality in low-resource settings

Challenges in ECE quality measurement are amplified in LMIC by the need to adapt measures to better match context and the constraints on resources for measurement (Wolf et al., 2018; Raikes, Sayre et al., 2019). Culturally and contextually sensitive measurement is important for long-term usefulness and applicability of results (Nonoyama-Tarumi, Loaiza, & Engle, 2009; Sabanathan, Wills, & Gladstone, 2015; Wuermli, Tubbs, Petersen, & Aber, 2015). Concerns have been raised that the underlying constructs of many quality measures may not be appropriate for all settings without adaptation. The ECERS-R, for example, is viewed by some as reflecting a "Western view on quality ECEC" (Hu, 2015), and several authors have voiced concerns about assuming one definition of quality applies everywhere (e.g., Myers, 2004; Urban, 2019). The risk of using measures from another context is the imposition of standards of child development and quality across diverse contexts,

which inadvertently reinforces inequity by marginalizing locally generated definitions of quality and child development (Dahlberg, Moss, & Pence, 1999; Myers, 2004; Urban, 2019).

Although the literature base outlining details of measure adaptation is not extensive, adaptations of quality measures have included changes at the domain, construct and item levels. Using a version of the MELE in Colombia, Ponguta et al. (2019) report that changes of domains, constructs and items were all required to build an appropriate ECE quality measure for a national preschool evaluation. Garvis, Sheridan, Williams, and Mellgren (2018) reported that the ECERS-3 had many indicators with good alignment to Swedish settings, but that some key constructs – such as Swedish views of the child's perspective – required item-level adaptation. Comparing the ECERS-R with a Chinese Kindergarten Quality Rating System (KQRS), Hu (2015) found that the ECERS-R concurred with many but not all elements of quality as defined by the national system. Of note was the misalignment between the ECERS-R and the KQRS on teacher/child interaction, attributed to the different philosophies of the Chinese and US systems (for example, whole-group instruction as a manifestation of Chinese collectivist philosophy). An Indian version of the ERS was developed, called the Early Childhood Education Quality Assessment Scale, containing several adaptations of items to better match Indian context, such as modification of items to address social inclusion (Kaul et al., 2014).

At the same time, studies conducted in sub-Saharan Africa and other low-income regions indicate some relevance of constructs derived from ERS and CLASS, especially evidence of external validity, or associations between the quality of ECE programs and child development (e.g., Jackson et al., 2019; Rao et al., 2017). Using a modification of the CLASS designed for low- and middle-income country settings called the Teacher Instructional Practices and Processes Scale (TIPPS; Seidman, Raza, Kim, & McCoy, 2014) within a large sample of preschools in Ghana, Wolf et al. (2018) found that the TIPPS yielded three reliable factors that predicted children's learning at the end of the school year with small to moderate effect sizes (f^2 <.20; see Cohen, 1988), after controlling for baseline scores. Using the same sample, McCoy and Wolf (2018) found that TIPPS scores at the start of the year predicted growth in children's learning over the course of the year. In a comprehensive look at quality of early childhood education across the developing world, Rao et al. (2017) found that studies in 40 countries provided evidence that early childhood intervention programs can have an impact on child cognitive development, and also noted variations in program quality but did not find significant associations between quality and intervention effect size in their analyses.

Evaluations of the Madrasa Early Childhood Development Program developed in Kenya and implemented in several East-African countries (Kenya, Zanzibar, Uganda) have also demonstrated positive program impacts of higher-quality settings, as measured by the ECERS-R and ECERS-E. Children in both government and Madrasa programs had higher overall scores at post-test than children who stayed home, and children attending Madrasa schools had higher scores than children attending government schools (Mwaura, Sylva, & Malmberg, 2008). In a follow-up study, children attending the Madrasa program saw greater gains over time as a function of quality and quality was more strongly associated with child outcomes in the Madrasa program (Malmberg et al., 2011). Evidence of associations between scores on quality measures and child development have also been reported in Bangladesh (Moore, Akhter, & Aboud, 2008), Indonesia (Brinkman et al., 2016), and Chile (Leyva et al., 2015). In a ten-country longitudinal study of child development and quality in preprimary settings, Montie, Xiang, and Schweinhart (2006) found children benefited from teachers with more years of education; time spent in free choice and non-whole group activities; and access to a variety of materials and equipment, regardless of country of residence, reporting small to moderate effect sizes across all significant effects.

Little work has addressed internal structures of ECE quality scales in LMIC. Using both confirmatory and exploratory factor analyses, Wolf et al. (2018) report on three distinct factors from the TIPPS focused on teachers' interactions with children. Two of these factors showed associations with child learning, although only three of 12 tested associations between quality factors and child learning were statistically significant. Using exploratory factor analyses of the Global Guidelines Assessment, which is a self-assessment tool, Hardin, Bergen, and Hung (2013) reported an overarching "quality" factor that included items from all five of the proposed subscales, along with other factors that reflected groups of items that were roughly analogous to designated subscales, but did not test associations with child development. Therefore, while existing work has identified a myriad of different structures, a basic formulation comprised of teacher/child interaction; access to learning materials; and health/safety may adequately capture the fundamental aspects of quality in LMIC, drawing on work by Gordon et al. (2013); Early et al. (2018), and Mariano et al. (2019) documenting distinct factors indexing teacher/child interactions and language environments, physical characteristics of settings, and exposure to learning activities.

Against the backdrop of the mixed psychometric evidence on existing measures and the needs for contextually sensitive and feasible measurement, MELQO's Consortium developed a new tool, the MELE. Priority was placed on developing an approach that reflected empirical work on child development but could be aligned to local standards, especially given the concern that tools from other contexts could contribute to inequity by imposing foreign quality standards that were not in line with national goals and cultural priorities (e.g., Myers, 2004; Urban, 2019). A free and open-source tool to reduce measurement costs in resource-constrained systems (see Aboud & Proulx, 2019) was also envisioned.

The "core" MELE (MELE-C) that served as the starting point for country adaptation was drawn from existing measures of quality including the CLASS, ERS, and the Global Guidelines (Sandell, Hardin, & Wortham, 2010; see UNESCO et al., 2017, for a detailed description). MELE outlined seven "dimensions" and related items that were deemed by UNESCO's consortium as potentially relevant to quality across settings, with guidelines for adaptation that prioritized convening local stakeholders and aligning items to national standards. The MELE-C scale, prior to country adaptation, shares many elements with other measures, including a focus on teacher/child interactions; materials; and physical environments, such as health and safety (see ecdmeasure.org for a recent version of MELE). Released in 2016, the MELE-C has now been used in several countries. For example, in Malawi, Shallwani, Abubakar, and Kachama (2018) describe quality in community-based child-care settings using a version of MELE. In Indonesia, Proulx and Aboud (2019) used a version of MELE to detect intervention effects in preschool quality due to disaster risk reduction initiatives. In Colombia, as noted above, Ponguta et al. (2019) describe the adaptation process of MELE to adhere to local quality standards.

This study describes the adaptation and resulting psychometric properties of MELE scores in one sub-Saharan African country (denoted as MELE-A). First, we hypothesized that the MELE-C constructs and items would be generally aligned with the content of national standards, in line with growing adoption of global standards into national curricula (Raikes, Davis et al., 2019), but that adaptations would also be required to improve contextual fit to create a revised version, MELE-A. Second, we hypothesized that the MELE-A items would factor into three correlated but distinct constructs: health/safety, materials/activities, and teacher–child interactions. Third, we hypothesized that MELE-A scores would demonstrate adequate internal consistency (reliability) evidence.

Finally, we hypothesized that MELE-A scores would be associated with teacher and school characteristics, and positively predict child development (also using an adapted tool; see (Raikes, Koziol, Janus, Platas, Weatherholt, Smeby & Sayre, 2019), in line with theory on child development and quality, and building on existing studies of quality and teacher characteristics.

4. Method

The MELQO initiative supports low- and middle-income countries in generating feasible, actionable measurement of early childhood development and quality of preprimary settings, to inform both global monitoring and provide nationally relevant data (Raikes, Sayre et al., 2019). The government of the focal country in this study agreed to join the MELQO project to generate information on quality of preprimary education. The SSA country has a population of over 2 million people and close to half are children under the age of 17. The Human Development Index (a measure of social and economic dimensions) ranks the social and economic conditions in this country among the bottom quarter of all countries (United Nations Development Program, 2019). Approximately 70 percent of the population lives in rural areas with limited access to basic health and education services, and most of the population lives below the international poverty line of US $1.90 per person per day. About 33 percent of children are enrolled in preprimary education, and enrollment has more than doubled since 2000. Early childhood education is delivered through public (government-run and financed) and private (privately ownership, unregulated, fee-based for parents); most programs are private and in urban areas and operate full days, while government-run preprimary schools typically operate for half-days.

The country has made several advances in early childhood education in recent years, with the adoption of a national policy for integrated early childhood development and care, followed by a strategic plan and inclusion of ECE in the education sector planning process. The role that these documents played in the adaptation process is outlined in greater detail below. The MELQO team, consisting of a technical advisor, members of the ministry of education and other stakeholders, participated in an adaption process outlined in detail in the procedures section, while the government took responsibility for all activities related to data collection and protection of human subjects.

4.1. Participants

A sample of 250 schools (85% private, which reflected the proportion of private to public schools across the country) across four country zones (defined by regional ecology and representing all regions of the country) were recruited for participation in the MELQO study based on government lists. Some private preprimary facilities were not included in government lists because they were not known or registered with the government, so sampling should not be considered comprehensive. Children were nested in classrooms (one classroom per school), with one main teacher serving as the focal point of the observation and providing information on children's social/emotional development. From each classroom, approximately 4 children were randomly selected to complete measures of child development and learning for a total sample size of 979 children (46% male; mean age = 5.49 years [SD = 0.54]). Classroom observations, which were used to complete the MELE-A, lasted on average 2.02 h (SD = 0.71) and took place in the morning during a scheduled visit. On average, teachers reported 9.10 years of preprimary experience (SD = 7.31) and nearly half (49%) received professional development in the past 12 months. The average classroom had 19.94 children (ranging from 2 to 200; SD = 17.12). Additional child and family characteristics of the sample are presented in Table 1 and show that on average families had fewer than half of the measured assets (average proportion of total assets listed = .43; assets included electricity, radio, television, landline phone, refrigerator, heating, cooling, running water, gas stove, paraffin stove, livestock, mobile phone, bicycle, car/truck, animal-drawn cart). Parents reported 13% of children had a disability and most mothers' (59.5%) highest level of education was primary school or below. Additional classroom/teacher and school characteristics are presented in Table 2. On average, teachers were just under 40 years of age, relatively satisfied with their job (mean

Table 1
Child and family characteristics.

	N	M	SD
Child age in years	860	5.49	0.54
Family assets	903	0.43	0.18
	N	N	%
Child is male	906	420	46.4%
Child has disability	900	117	13.0%
Mother education level	677		
Less than primary school		88	13.0%
Primary school completion		315	46.5%
Junior certificate		126	18.6%
Cambridge oversees school certificate/general certificate of secondary education/form E		79	11.7%
Certificate/diploma		48	7.1%
1st degree or above		21	3.1%

Note. N = total sample size with information on variable. n = group size. assets = proportion of assets out of 15 (electricity, radio, television, landline phone, refrigerator, heating, cooling, running water, gas stove, paraffin stove, livestock, mobile phone, bicycle, car/truck, animal-drawn cart).

Table 2
Classroom and school characteristics.

	N	M	SD
Teacher age in years	222	39.66	12.06
Teacher years of preprimary experience	208	9.10	7.31
Teacher perceptions			
Satisfied with job	222	4.07	1.08
Receives adequate support from head teacher/school board	221	3.98	1.18
Overwhelmed by amount of work	222	2.28	1.22
Has adequate resources to carry out teaching duties	222	2.36	1.31
Has training needed to be an effective pre-primary teacher	222	3.35	1.34
Classroom size (number of children)	242	19.94	17.12
Teacher–child ratio	243	0.11	0.07
	N	n	%
Teacher is male	221	26	11.8
Teacher education level	221		
Less than primary school		9	4.1
Primary school completion		54	24.4
Junior certificate		63	28.5
Cambridge oversees school certificate/general certificate of secondary education/form E		41	18.6
Certificate/diploma		46	20.8
1st degree or above		8	3.6
Teacher received professional development in past 12 months	220	108	49.1
School is private	247	209	84.6
Ecological zone	250		
Zone 1		69	27.6
Zone 2		136	54.4
Zone 3		18	7.2
Zone 4		27	10.8

Note. N = total sample size with information on variable. n = group size. Teacher perceptions rated on a 5-point scale ranging from 1 = strongly disagree to 5 = strongly agree.

of 4.07 on a 5-point scale with 5 being high), with variable levels of education (about one-fourth with primary education only, about one-third middle-school level, one-fifth a high school education and another one-fourth with some college or higher), and 11.8% of teachers were male. Because this study was considered an observation of routine educational practice and was administered by the government for planning purposes, data collection was exempt from IRB oversight. This study was reviewed and approved by the first author's institution as exempt under category 4b at 45 CFR 46.104 under the 2018 Requirements. Only deidentified data were analyzed as part of this study.

4.2. Measures

To create measurement tools, MELQO's Consortium, comprised of individuals representing academia, non-profit organizations, multi-lateral organizations and governments, defined a common item set drawn from existing measures used in low- and middle-income countries that represented constructs deemed by a large international consortium to be potentially relevant across countries. MELQO also developed procedures for adaptation used by low- and middle-income countries to adapt and measure child development (see UNESCO et al., 2017, for a complete description of MELQO). Two sets of tools were created: the MODEL consists of a child direct assessment, and reports on children's development completed by teachers and parents; the MELE consists of a classroom observation measure, a teacher interview, and a director/head teacher interview (UNESCO et al., 2017).

4.2.1. Classroom observation tool

The adapted version of the Measuring Early Learning Environments (MELE-A) scale was used to observe classroom quality. The adaptation process began with the core MELE-C and was focused on two primary goals: first, ensuring alignment with policy documents and cultural priorities; and second, ensuring the feasibility of the measure, including the ability of observers to use the tool reliably. There were three phases of the MELE adaptation process: (1) a review of existing policy documents related to ECE; (2) a stakeholder workshop comprised of government officials, civil society, researchers and other ECE stakeholders to generate a draft version of MELE-A; and (3) small-scale pilot testing followed by revisions to the MELE-A to create a final version.

4.2.1.1. Phase 1: Review of current country ECE landscape. This phase began with the MELE "core" and focused on its alignment with key policy documents, including the national policy and strategic plan, the education sector strategic plan, and the recently developed standards addressing child development and learning. The alignment of these policies to the "core" MELE is found in Table 1S.

4.2.1.2. Phase 2: Stakeholder adaptation workshop. After completing Phase 1, the government organized a one-week in-person adaptation workshop to discuss the adaptations required to the "core" MELE tool. An intersectoral group of attendees representing ministries of education, health/nutrition, child protection, agriculture and food security, social development, statistics and partner organizations including staff from training institutions, universities, NGOs, and bilateral organizations attended the workshop. Participants reviewed Phase 1 alignment and conducted an item-by-item review of tools to determine appropriateness or need for revisions. During the adaptation workshop, workshop participants visited four ECE programs (one private, two public and one community-based) to observe classroom practices, collect video of lessons and interactions, and take photographs of space and materials. Small focus groups with parents were conducted to gain their feedback on the tools.

Several changes were made to the tools as a result of the workshop discussion and visits. For example, observations of the schools informed the list of materials included in the descriptors for the educational and fine motor materials items; the content of the tools was modified to make the tool more relevant to curricula and teacher training by retaining an item on teachers' use of a "theme" in teaching. A point of misalignment was the MELE-C's lack of focus on plants and animals, which was part of both the curricula and child development and learning standards. Attendees determined that a separate item on plants/animals was not needed, rather examples related to use of natural materials were included in the scoring rubric for other items (such as fine motor activities). This process resulted in MELE-A and MODEL tools that maintained both a core set of items but also reflected country-specific priorities.

4.2.1.3. Feasibility was also addressed. Items related to classroom arrangement were revised to be scored as yes or no versus on a 4-point scale to simplify items. Items on gender equity in the classroom and teachers' use of discipline tactics were removed to reduce the number of items and because they were not considered high priority for the country stakeholders and viewed as challenging for reliability. For example, the item on discipline asked the observer to score how well the teacher disciplined the children when a misbehavior occurred (from not at all to redirecting behavior and helping children understand the reason for rules). Stakeholders expressed that child misbehavior was a very rare occurrence in classrooms and observers could interpret the same behavior differently (for example, tone of voice, nonverbal language and even instances in which children should or shouldn't be disciplined), so the item was deleted. Decisions to shorten the total number of items were also based on cost considerations, to avoid extending the observation time, number of data collectors required or length of training time.

4.2.1.4. Phase 3: Small-scale pilot-testing & preliminary training. Following adaptation, the tools were translated, reviewed, and piloted. Ministry staff were trained on administering the tools through webinars with the MELQO team and additional practice. The study team collected data using paper forms from preprimary programs in teams over a two-week period. The study team also collected additional data (children's work samples, feedback from participants, reviews of translations, documentation of issues) and reported out findings related to tool revisions needed, preparation for training, and procedural challenges. Pilot data were collected from 38 schools that were randomly selected from a list of known preprimary schools across four ecological zones. The overall purpose of the pilot was to test the process of collecting these data, ensure that the adaptations of the tools were appropriate, identify problematic items that were redundant or hard to understand, and build capacity within the team for the national study data collection. The data collection team also identified issues with tool translation, documented the time it took to administer each tool, noted challenges related to scheduling and coordinating data collection, and collected video of classroom activities and samples of children's responses to items to be used for the national study training.

As a result of the pilot, further revisions were made to clarify items and prepare for training. For example, the child engagement item, focused on whether children were actively engaged in activities during the observation, was originally removed from the pilot version but added back in for the national study due to feedback from observers that the item was valuable, and revisions to the toilet facilities rubric were made. The national study sample size was reduced because of the high transportation costs discovered during piloting. A description of all adaptations and the rationale for making the adaptations is in Table 1S in the Supplementary material. The complete measure used appears in the Supplement.

The MELE-A was completed by trained observers over a 2-h morning observation period. Teachers were aware of the visit and directed to conduct the class as they normally would. Observers were in the classrooms and followed the teacher and children if they went outside or to other areas. The observers scored all items in this one session. Curricular activities, health and safety, and teacher interaction items were score on a 4-point rubric while materials/activities and materials items were scored as yes/no. Item descriptive statistics are provided in Table 2S in the Supplementary material.

4.2.2. Child development and learning and social/emotional development

The MODEL tools were used to assess child development and learning and social/emotional development. The MODEL has two components: First, parents and teachers were interviewed on the social/emotional development of each child. Parent-report data were available for 906 children. Each participating teacher reported on a random subset of the participating children in his/her classroom to reduce the data collection burden on teachers (M = 1.99 students/teacher [SD = .15], range = 1–3 students), such that teacher-report data were available for 473 children. Table 3S in the Supplementary material lists the ten 3-point Likert-type items and descriptive statistics based on both parent and teacher report. For all items, all response categories were used suggesting variability in social/emotional development. In general, both parent and teacher perceptions were generally positive.

Second, children's development and learning was assessed by a trained assessor with a battery of tasks relating to spatial vocabulary, verbal counting, producing a set, number identification, letter identification, expressive language, listening comprehension, name writing, head toes knees shoulders, and pencil tap. This direct assessment (DA) took on average about 30 min to complete with each child. As with the MELE, the MODEL was adapted and aligned with national standards (Raikes, Koziol et al., 2019), for a detailed description of MODEL and psychometric properties from one country). See Table 4S for item descriptive statistics. Most items were scored dichotomously except for verbal counting, name writing, and head toes knees shoulders. Item difficulty ranged widely across tasks, with letter identification items tending to be very difficult and expressive language very easy.

4.2.3. Child and family characteristics

In addition to reporting on their child's social/emotional development, parents were interviewed by trained data collectors about demographic and background information such as child age, gender, and disability status, family assets, and mother education level (see Table 1).

4.2.4. Teacher characteristics

Teachers were interviewed by trained data collectors on demographic and teaching characteristics, including their age, gender, education level, years of preprimary experience, professional development experiences in the past year, and perceptions of their job (see Table 2).

4.3. Procedures

4.3.1. Training

Observers were trained to reliability on MELE-A and MODEL measures that had been adapted to national context and revised from the pilot phase. All enumerators for the MELQO tools were recruited and hired by the Ministry and included ministry staff, national teacher trainers, students and graduates from the national university. For the MELE-A, six of the enumerators were Ministry staff, two worked with ECCD programs and two were students from a statistics program. The six-ministry staff had previously been trained on the MELE-A for the pilot. Training was implemented in-person by members of the global MELQO team and lasted five days per measure, using presentations, the tools, manuals, activities, video clips, role playing and visits to schools to practice administering instruments. The site visit was used to familiarize observers with the logistics of completing observations and to talk informally about what they were seeing and how it might be scored. During the visits, enumerators training on the survey tools practiced interviewing teachers. Due to scheduling and transportation challenges, live inter-rater reliability visits for the MELE-A were not possible. As part of training and in preparation for reliability, the MELE-A observers watched videos and scored them independently and then scores were compared and discussions around clarifying scores took place. For items that were more difficult for observers, additional time was spent reviewing those items. Two reliability tasks were used to certify observers, answering all items on a quiz correctly and scoring 85% of items as an exact match to a consensus score using a video; all data collectors used for the study met this standard before collecting data (and only one person did not meet this standard for the MELE-A and therefore did not collect data). It was not possible to collect inter-rater reliability or data on observer drift in the field due to costs, and teams consisted of only one MELE-A observer each and were assigned to specific regions. At the end of the training, an addendum to the training manual that covered clarifications that arose during the training sessions was provided to all observers. To ensure quality during the data collection period, supervisors were available to respond to questions and checked protocols daily.

4.3.2. Data collection

Independent teams of data collectors were used to conduct classroom observations and child assessments, so that all classroom observations were blind to the child assessments. The teams consisted of one person responsible for the classroom observation, one person responsible for the child assessment, and two people responsible for the survey data collection. Schools and teachers were notified of the data collection day and procedures. On average, classroom observations lasted 2.02 h (SD = .71). Parents were informed of the data collection date and asked to be present at the school that day. After children were randomly selected, parents of selected children were asked to stay to complete interviews or were contacted later to complete interviews; some parents did not participate in the interviews. Child assessments and parent interviews occurred during or after the classroom observation. Ninety-three percent of all parents completed surveys on child development. Data collectors completed teacher interviews after conducting classroom observations; complete data was obtained for 89% of all teachers.

4.4. Data analyses Plan

For this study, we selected three key constructs that have been shown to influence child development in other studies: health/safety, materials/activities, and teacher–child interaction. Health/safety refers to the conditions of the school potentially adverse for children's health and safety: access to clean water and toilets and exposure to dangerous facilities. Materials/activities refers to the range of activities (both curricular and play) children experienced during the observation, including the number and type of materials they engaged with. Teacher/child interaction refers to multiple aspects of teachers' interactions with children, including use of playful learning, engagement and negativity with children, and teachers' use of individualized instruction. See Fig. 1 for a summary. The 4-point scoring rubrics for health and safety items (drinking water, handwashing, toilet facilities and safety

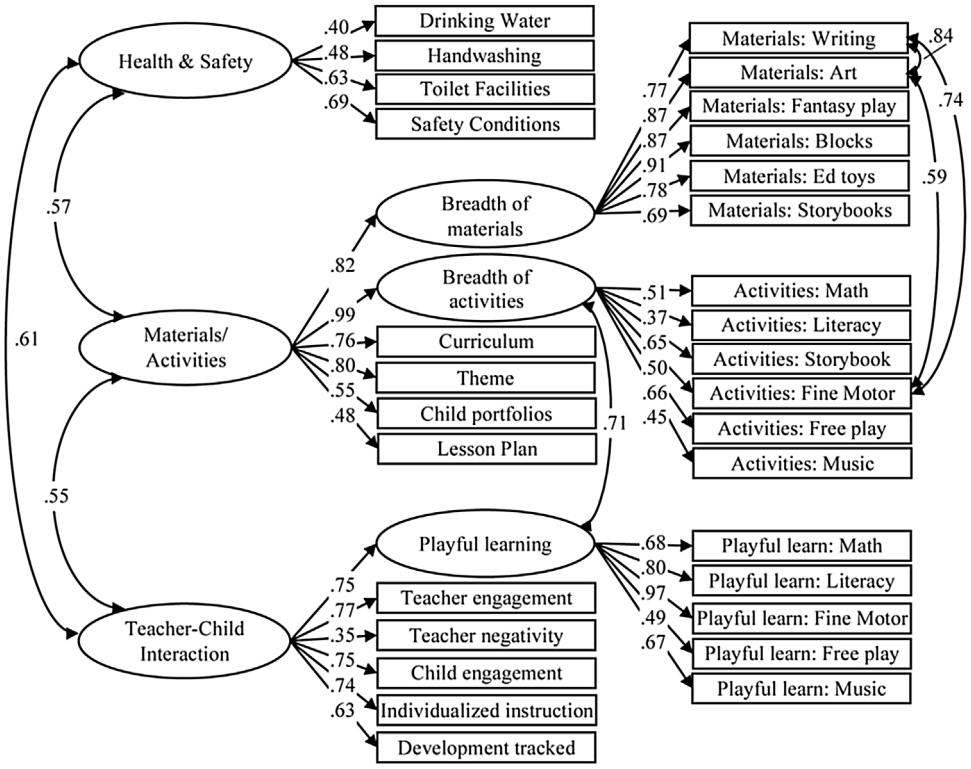

Fig. 1. Simplified path diagram of classroom quality construct. Numeric values are standardized coefficients.

conditions) described various levels of these conditions, for example, more sanitary drinking or handwashing conditions. Materials were coded as available, meaning present in the classroom, or not available. To disentangle curricular content from teacher–child interaction, each of six learning activity items (math, literacy, storybook, fine motor, free play, music/movement) was recoded into two separate items: (1) whether or not the activity occurred during the classroom observation (item labeled "activity" and scored as observed or not observed); and (2) the extent to which the activity embodied playful learning principles, including pedagogy of academic material with games, songs, toys, pretend play and dialogue with children (item labeled "playful learning"). If the activity was not observed, then the corresponding playful learning item was treated as missing. Playful learning was defined by the rubric as teaching practices that included children engaging with materials, exploration, free or open play, choice or discussion (see Table 2S). As further described in the data analysis plan, a two-part modeling approach in conjunction with the use of all available data ensured that cases with missing data on these variables were still included in the analysis and that the reason for the missingness (i.e., no activity observed) was reflected in the model. Storybook reading was observed in only 18% of classrooms, so the corresponding playful learning item was omitted from the model due to inadequate covariance coverage.

Quantitative analyses were designed to evaluate three of the four sources of reliability and validity evidence we aimed to assess: Evidence based on internal structure, assessed via categorical confirmatory factor analysis (CCFA); evidence of internal consistency, assessed via reliability coefficients; and evidence based on relations to other variables, assessed via regression. Evidence based on test content, assessed through scale development and country reports of the degree of alignment between standards and MELE-A items, did not rely upon quantitative analyses, and was assessed by examining the degree of alignment with government standards as outlined in the Results section.

4.4.1. Evidence of internal structure

CCFA, assuming ordinality of response options, was performed to evaluate the hypothesized 3-factor structure of the classroom quality scores. See Fig. 1 for a simplified path diagram of the factor model. The three ovals on the left-hand side represent correlated but distinct quality constructs. Each construct is measured by multiple items represented by rectangles. Note that three sets of items pertaining to materials, activities, and playful learning are presumed to measure lower-order constructs (breadth of materials, breadth of activities, and playful learning), where these lower-order constructs are then presumed to be indicators of the target quality constructs. As mentioned in the Measures and Variables section, a two-part modeling strategy in which learning activity items were modeled via two separate constructs (breadth of activities and playful learning) allowed for clearer differentiation between quality constructs. A correlation was specified between the breadth of activities and playful learning constructs given that missingness on the playful learning items was directly tied to whether the corresponding activity was observed. Additional residual correlations were specified between individual items where empirically and conceptually justified.

The aim of this study was to evaluate the psychometric properties of the classroom quality scores, but CCFA was also performed on the child outcome items in order to derive scores for use as criterion variables. For the child DA, a higher-order factor model was fit to the data based on prior research that identified strongly correlated domains pertaining to children's early learning and development (cf., (Raikes, Koziol et al., 2019)). See Fig. 1S in the Supplementary material for a simplified path diagram. The individual DA items were specified to load on one of three lower-order domain factors (early mathematics, language/literacy, or executive function), and the three domain factors were specified to load on one higher-order early learning and development factor. A bifactor parameterization was used to account for dependency among items within the same task such that each item also loaded on a task-specific

factor. The task-specific factors were assumed to be uncorrelated with each other (with exceptions noted in the Results section) and uncorrelated with the domain factors. For both the parent- and teacher-reported social/emotional development measures, a single-factor model was fit to the data, with residual correlations allowed where empirically and conceptually justified.

Analyses were performed in M*plus* Version 8 (Muthén & Muthén, 1998-2017) using the weighted least squares mean- and variance-adjusted (WLSMV) estimator with theta parameterization and a probit link function. All cases with at least partial data were included in the analyses. See Asparouhov and Muthén (2010) for a more detailed discussion of how missing data is handled with WLMSV estimation. The primary latent factors were identified by fixing the mean to 0 and variance to 1, with the lower-order factors of the child DA models identified by fixing a referent item threshold to 0 and factor loading to 1. For the child DA and social/emotional development CCFAs, Taylor series linearization variance estimation (via M*plus* TYPE = COMPLEX and specification of cluster variable) was used to compute clustered standard errors to account for nesting of children within classrooms.

Global model fit statistics/indices and local fit information (modification indices, pattern of standardized loadings) were used to evaluate evidence regarding the internal structure. Evidence of adequate global fit is suggested by a non-significant chi-square test of exact fit; root mean square error of approximation (RMSEA) $\leq$.08 or .05 (Browne & Cudeck, 1993); and comparative fit index (CFI) $\geq$.90 (Bentler, 1990) or .95 (Hu & Bentler, 1999); although strict adherence to cutoffs is not recommended (Kline, 2016).

4.4.2. Evidence of internal consistency

Model-based coefficient ω was used as an average measure of internal consistency (reliability). In the context of CCFA, reliability of scores varies across the latent variable continuum, where total information functions (TIFs) are typically used to illustrate this variability. However, when factors are measured by both continuous and categorical indicators (as is the case here), the TIFs provided by M*plus* underestimate the true reliability because information from the continuous indicators is not captured in the functions. Construction of the TIFs requires numerical integration, so it is not possible to construct the functions manually. We acknowledge this unavoidable limitation and recognize that providing an average measure of internal consistency oversimplifies reliability evidence.

4.4.3. Evidence of relations to other variables

Validity evidence based on relation to other variables was evaluated by estimating (a) associations between teacher/classroom/school characteristics and classroom quality constructs (convergent evidence), and (b) concurrent test-criterion associations between classroom quality constructs on child outcomes (DA early learning and development scores and parent- and teacher-reported social/emotional development scores). All teacher/classroom and school characteristics presented in Table 2 were considered as potential predictors in the first set of associations. Because many of these variables are highly correlated which may result in multicollinearity if all variables are considered simultaneously, we used a stepwise approach in which only significant predictors were retained in the final models. For the second set of associations, we controlled for child age, child gender, family assets, mother education, teacher years of preprimary experience, teacher education, teacher perception of adequacy of resources, and ecological zone. Other child/family, teacher/classroom, and school characteristics listed in Tables 1 and 2 were omitted as covariates because they were not uniquely associated with any child outcomes after accounting for the other covariates.

Maximum a posteriori (MAP) estimation was used to derive quality and child outcome factor scores based on the CCFA models. To account for unreliability in the factor scores, the scores were treated as single-item latent factors with residual variances constrained to be equal to the variance of the factor scores multiplied by one minus the estimated reliability (i.e., coefficient ω) of the scores (Brown, 2006; Raikes, Koziol et al., 2019). Analyses were performed in M*plus*, with robust maximum likelihood estimation (MLR). For child outcome analyses, Taylor series linearization variance estimation was used to compute clustered standard errors. Statistical significance was evaluated based on α = .05. Practical significance of individual variables was assessed via standardized regression coefficients (the standard deviation unit change in the outcome associated with a one standard deviation unit increase in the predictor) and Cohen's f^2 (the unique proportion of variability in the outcome accounted for by the predictor), where f^2 = .02, .15, and .35 are guidelines for defining small, medium, and large effects, respectively (Cohen, 1988). In addition, R^2 was calculated to determine the total proportion of variability in the outcome accounted for by the complete set of predictors.

5. Results

5.1. Descriptive statistics

Table 2S lists the MELE-A items, response options, and descriptive statistics. Approximately 60% of classrooms had access to a sanitary water source, 23.1% had running water or hand poured system and soap available for handwashing, 3.7% had flush or pour-flush toilets, and 19.2% had no dangerous conditions on school grounds or in the classroom. Various materials (writing utensils, art, fantasy play, blocks, educational toys, storybooks) were available in more than 50% of the classrooms. Likewise, various activities (math, literacy, fine motor, free play, music/movement) were observed in most classrooms, although storybook reading was observed in only 17.8% of classrooms. More than 60% of the classrooms followed a curriculum and used a lesson plan and theme for organizing activities, whereas less than 30% of classrooms used child portfolios and tracked children's development on a regular basis. Most teachers never (73.6%) or rarely (17%) exhibited negativity, and in most classrooms all (29.1%) or most (43.6%) children were engaged most of the time. Across items, all response options were used. Items with small frequency counts/imbalances across response options included drinking water, toilet facilities, math activities, teacher engagement and negativity, and child engagement.

5.2. Evidence of test content

Content validity evidence of the MELE-A scores is demonstrated in two ways: First, by the scale development, in which items were created by drawing upon the literature on quality indicators, existing quality measures and an expert review panel (UNESCO et al., 2017), and second, during the country adaptation process, outlined above. To assess test content validity, an alignment document with each of the core MELE items and content of government policy documents was created. The national policy described the priority placed on ensuring quality in ECE, specifically in three areas: (1) improving levels of quality in preprimary programs and, in particular increasing access to high-quality programs for children living in poverty, (2) addressing inadequate pre- and in-service training for teachers, and (3) establishing a system of quality assurances and accountability of ECE services. However, policy documents did not include specific, measurable indicators of quality in ECE, beyond tracking teacher attendance and state of school buildings in public schools only. Because of the limited information on definitions of quality, items indexing child development and learning (from the ELDS) were included in the alignment tables to identify if and how MELE measured teaching practices or classroom characteristics to

Table 3
Internal consistency and global model fit.

	N	Coef ω	χ^2	df	p	RMSEA	RMSEA 90% CI	p (RMSEA $\leq$ .05)	CFI
i. Classroom quality[a]	247		715.72	398	<.001	.057	[.050, .064]	.047	.896
ii. Classroom quality[b]	247		688.92	395	<.001	.055	[.048, .062]	.118	.904
Health & safety		.65							
Materials/activities		.87							
Teacher–child interaction		.83							
iii. Classroom quality[c]	247		698.823	395	<.001	.056	[.049, .063]	.080	.899

Note.
[a] No post-hoc residual correlations.
[b] Final model with residual correlations between (1) Materials: Writing Utensils and Materials: Art; (2) Materials: Writing Utensils and Activities: Fine Motor; and (3) Materials: Art and Activities: Fine Motor.
[c] Model with response options 3 and 4 collapsed for drinking water item, and response options 1 and 2 collapsed for individualized instruction item.

give children opportunities to gain ELDS skills. For example, the national policy stated that ECE programs should be holistic and include "perceptual, language, cognitive, physical (gross and fine motor), social and emotional development, including the ability to regulate their behavior" and the ELDS included indicators of child development in these areas. If MELE items addressed implementation of math, literacy/language, fine motor, and gross motor activities, these items were viewed as "aligned" with government policy. The government documents also stated that children would have access to sanitary water; while access to clean water and sanitation were included in the "core" MELE, the appropriate indicators of access to water required local adaptation, such as the use of wells, water carts, bottled water, etc., which was undertaken in the next phase.

Because the government documents lacked specific guidance on teaching practices or materials, an estimate of direct alignment of MELE with government policy documents could not be made. The ELDS were very specific and included 43 standards with over 150 indicators for children ages 36–60 months, which in turn indicated if MELE items included elements of classroom environments necessary for children to learn skills outlined in the ELDS. For the 22 ELDS subdomains, there were related MELE items for 13 (or almost 60%) primarily focused on language, literacy and mathematics. Some areas were not well represented; for example, there were no MELE items related to ELDS domains Cultural Heritage and Life Skills. A similar adaptation process was also completed for the MODEL and integrated into a crosswalk document.

5.3. Evidence of internal structure

Model fit information is provided in Table 3 for the classroom quality CCFAs, and Table 5S in the Supplementary material for the child outcome CCFAs. Complete measurement model parameter estimates are available in the Supplementary material (see Tables 6S–9S).

For classroom quality, residual correlations were allowed for three pairs of items: (1) Materials: Writing Utensils and Materials: Art; (2) Materials: Writing Utensils and Activities: Fine Motor; and 3) Materials: Art and Activities: Fine Motor. These adjustments were made post-hoc based on empirical evidence (modification indices) and conceptual justification, that is, based on the fact that examples of writing utensils and art materials overlapped (e.g., pencils, pens, chalk), and examples of fine motor activities included activities that made use of writing utensils and art materials (e.g., writing/scribbling, drawing). Whereas the three classroom quality factors were strongly correlated (r = .57 [health/safety and materials/activities], .61 [health/safety and teacher–child interaction], and .55 [materials/activities and teacher–child interaction]), the hypothesized 3-factor structure fit the data significantly better than a unidimensional structure ($\Delta\chi^2$[df = 3] = 73.86, p < .001) suggesting that the factors are distinct. Although the test of exact fit (χ^2) was rejected for the 3-factor model, the test of close fit based on RMSEA was not rejected and CFI was acceptable, lending support to the hypothesized internal structure.

All items loaded strongly on their respective factors (all standardized loadings >.3). For items with only two response options, the positive loadings provide support for the assumption that higher response options are associated with higher quality. To evaluate the assumption of ordinality for items with more than two response options, separate factor score means were calculated for classrooms in each response category to evaluate whether the means monotonically increased across categories. Of the 13 items with more than two response categories, 2 did not exhibit strict ordinality. For the handwashing item, the mean health/safety scores were −.44 (no water available), −.23 (unprotected dug well/spring, rainwater, surface water), .39 (cart with small tank/drum, tanker truck, protected spring), and .25 (sanitary water source). The third response category was observed for only eight classrooms, which may have contributed to the reversal between the third and fourth categories. For the teacher engagement item, the mean teacher–child interaction scores were −1.07 (teacher seems irritated towards children. . .), −1.08 (teacher appears distracted or uninterested in children. . .), −.09 (teacher appears to enjoy some tasks or children. . .), and .90 (teacher genuinely appears to enjoy teaching. . .), suggesting that the first and second response categories are similar with respect to quality. To evaluate the sensitivity of the results to these reversals, a separate model was estimated in which the non-ordered response categories were collapsed. The original model and model with collapsed categories are not nested, but in evaluating model fit, the model with collapsed categories had less support. Factor scores were very highly correlated across the two models (.999, 1.000, and .996 for health/safety, materials/activities, and teacher–child interaction, respectively). For these reasons, and because inferences based on this post-hoc evaluation of ordinality may be susceptible to sampling error, the original model without collapsing categories was chosen as the final model.

Similar evidence was observed for the child outcome models. For the child DA, a residual correlation was deemed necessary between the number and letter identification tasks. This same dependency was observed in previous work (Raikes, Koziol et al., 2019), and may be due in part to common method variance as both tasks required the child to point to a symbol (number or letter) and then verbally identify the symbol. The hypothesized hierarchical factor structure fit the data significantly better than a simple unidimensional structure ($\Delta\chi^2$[df = 3] = 52.18, p < .001). The hierarchical structure was further preferred to a simple 3-factor structure because of the very strong correlations among the lower-order domains (r = .96 [math with lang/lit], .85 [math with executive function], and .83 [lang/lit with executive function]). As before, the test of exact fit was rejected, but RMSEA provided evidence of close fit and CFI was acceptable.

Standardized loadings on the higher-order factor and three lower-order factors were positive and exceeded .3 except for

Table 4
Standardized regression coefficients and 95% confidence intervals indicating the unique association of classroom/school characteristics with classroom quality scores.

	Health & safety	Materials/activities	Teacher–child interaction
Teacher age in years[a]	—	—	−.16 [−.31, −.02]
Teacher years of preprimary experience[a]	—	.24 [.11, .36]	—
Teacher perceptions[b]			
Satisfied with job	—	—	—
Adequate support school	—	—	—
Overwhelmed by amount of work	−.20 [−.37, −.02]	−.16 [−.30, −.02]	—
Adequate resources to teach	—	—	—
Has training to be effective teacher	—	—	—
Classroom size (number of children)[b]	—	—	—
Teacher–child ratio[b]	—	—	—
Teacher is male[c]	—	—	−.18 [−.32, −.05]
Teacher education level[d]	.29 [.13, .46]	.37 [.24, .50]	.27 [.13, .42]
Teacher prof. development in past 12 mo.[c]	—	—	—
School is private[d]	−.14 [−.28, .00]	—	—
Ecological zone[e]			
Zone 1	—	—	—
Zone 2	—	—	—
Zone 3	—	—	—

Note. N = 250. Non-significant predictors (indicated by —) omitted from the model. Female, no professional development, public school, and zone 4 = reference groups.
[a] Continuous.
[b] Ordinal with possible range of 1 = strongly disagree to 5 = strongly agree.
[c] Dichtomous.
[d] Ordinal with possible range from 1 = less than PSLE to 6 = 1st degree or above.
[e] Nominal.

the expressive language items. Most items with more than two response options demonstrated evidence of ordinality with the exception of head, toes, knees, shoulders items 1, 3, 8, and 13 (higher-order factor score mean slightly higher for self-corrects option than correct option), and verbal counting (factor score mean slightly higher for highest number reached = 16–20 than 21–25, but only 2.5% of students scored in the latter category so the mean may not be stable). Overall fit for a model with these categories collapsed had similar support to the original model but with a slightly lower chi-square test value and slightly higher CFI value. The higher-order factor scores across the two models were correlated at 1.000. Because collapsing categories did not impact the relative ranking of factor scores, and because the ordered nature of the partial credit items could be theoretically justified (and empirically justified when considering that the majority of like items demonstrated ordinality), the original model without collapsing categories was chosen as the final model.

For the parent- and teacher-reported social/emotional development models, a residual correlation was specified post-hoc between two items that asked how often the child stops an activity when told to and how often the child follows instructions. This dependency is likely because stopping an activity when told to is an example of following instructions. For both parent- and teacher-report, the test of exact fit of a unidimensional factor structure was rejected, but RMSEA and CFI were acceptable providing adequate evidence to support the internal structure. All standardized loadings were positive and exceeded .3, and the assumption of ordinality was supported for all items.

Correlations among the three child outcome measures were small to moderate (r = .18 [DA development and parent-reported social/emotional], .34 [DA development and teacher-reported social/emotional], and .20 [parent-reported social/emotional and teacher-reported social/emotional].

5.4. Evidence of internal consistency

Coefficient ω was slightly below the often cited .70 internal consistency threshold for the health/safety construct (ω = .65), which is due in part to the fact that only four items were used to measure this construct. On the other hand, internal consistency evidence was strong for the materials/activities and teacher–child interaction constructs (ω = .87 and .83, respectively). Internal consistency evidence was likewise adequate for the three sets of child outcome scores (ω = .88 [child DA], .76 [parent report], and .83 [teacher report].

5.5. Evidence of relations to other variables

A summary of the unique associations of teacher/classroom and school characteristics with classroom quality scores is given in Table 4. See Tables 10S–12S in the Supplementary material for full model results. Lower teacher stress (feeling less overwhelmed by the amount of work), greater teacher education, and public school funding were significantly and uniquely positively associated with health/safety (f^2 = .05, .09, and .02, respectively, indicating small associations), with these three variables accounting for 18% of the variability in health/safety scores. Greater teacher years of preprimary experience, less teacher stress, and greater teacher education were significantly and uniquely positively associated with materials/activities (f^2 = .07, .03, and .17, respectively, indicating two small and one moderate association) and together accounted for 20% of the variability in materials/activities scores. Finally, younger teachers, female teachers, and greater teacher education were significantly and uniquely positively associated with teacher–child interaction (f^2 = .03, .04, and .08, respectively, indicating small associations) and together accounted for 14% of the variability in teacher–child interaction scores. The remaining and vast majority (33) associations were non-significant.

A summary of the unique associations of classroom quality constructs on child outcomes after controlling for child/family, teacher/classroom, and school characteristics is given in Table 5, and results for the full models are presented in Tables 13S–15S of the Supplementary material. Only one of the nine associations between classroom quality and child outcomes, after controlling for all covariates, was significant. Specifically, only materials/activities was uniquely associated with scores on the child DA (f^2 = .02, a small association), and none of the quality constructs were associated with parent- or teacher-reported social/emotional development. Taken together, the quality constructs and classroom covariates accounted for a greater proportion of variability in the child DA scores (R^2 = .27–.29) than the parent-reported (R^2 = .05)

Table 5
Standardized regression coefficients and 95% confidence intervals indicating the association of classroom quality scores on child outcomes controlling for child, family, classroom, and school characteristics.

	DA development & learning	Parent report soc/emot dev	Teacher report soc/emot dev
Health & safety			
Bivariate	.22 [.11, .33]	.00 [−.11, .11]	−.02 [−.15, .12]
With child/family covariates	.12 [.02, .21]	−.05 [-.15, .06]	−.08 [-.22, .06]
With class/school covariates	.15 [.03, .28]	−.05 [−.17, .06]	−.04 [−.18, .10]
With all covariates	.08 [−.04, .19]	−.07 [−.19, .04]	−.08 [−.23, .06]
Materials/activities			
Bivariate	.23 [.15, .32]	.07 [−.03, .17]	.00 [−.11, .11]
With child/family covariates	.15 [.08, .23]	.04 [−.06, .14]	−.05 [−.16, .06]
With class/school covariates	.20 [.09, .30]	.00 [−.10, .11]	.00 [−.11, .11]
With all covariates	.16 [.07, .25]	−.01 [−.11, .10]	−.03 [−.14, .09]
Teacher–child interaction			
Bivariate	.21 [.11, .31]	.02 [−.09, .13]	−.05 [−.16, .07]
With child/family covariates	.12 [.03, .20]	−.01 [−.12, .10]	−.09 [−.20, .03]
With class/school covariates	.17 [.06, .27]	−.01 [−.13, .10]	−.06 [−.17, .05]
With all covariates	.09 [−.01, .19]	−.03 [−.14, .08]	−.10 [−.21, .01]

Note. N = 979. Gray shading indicates $p < .05$. DA = direct assessment. Child/family covariates = child age, child gender, family assets, and mother education. Classroom/school covariates = teacher years of preprimary experience, teacher education, teacher perception of adequate resources, ecological zone. Taylor series linearization variance estimation was used to compute clustered standard errors to account for nesting of children within classrooms.

and teacher-reported (R^2 = .08–.09) social/emotional development scores.

6. Discussion

Measurement of ECE quality is an important component of building an equitable early childhood system. This study reported on an adaptation process and evaluated psychometric properties of scores from open-source, adaptable measures developed through the MELQO initiative. Results offered some support that an adaptable tool, modified for feasibility and to align with country context, could demonstrate evidence of validity when used in a LMIC. Using the MELE-A, three hypothesized constructs of ECE quality – health/safety; materials/activities; and teacher/child interactions – were supported by confirmatory analyses. Content of MELE-A was mostly considered consistent with national standards and cultural expectations of ECE quality by key stakeholders, and quality scores showed adequate internal consistency, albeit less so for health/safety. However, as noted below, the small and infrequent associations between quality scores and child development, while consistent with other studies, also raise several questions regarding the extent to which this quality observation tool measures key constructs that promote child development.

Two rationales for using MELE – feasibility and adaptability – were evaluated in this study. Experiences from this study and other literature suggest that adaptations based on feasibility may be necessary in low-income country contexts due to resource limitations, and even with such adaptations, quality measures can still demonstrate workable psychometric properties that seem to have similar strengths and limitations as more established ECE quality measures. Adaptability of measurement was another central rationale for developing MELE, especially to improve alignment with policy. Basic constructs of MELE-C were viewed as generally applicable with country standards, but many changes were made to create MELE-A, demonstrating that item-level adaptations may be required to gain buy-in and ensure that measures capture conditions appropriately. The lessons from the adaptation process include the value of including a wide range of stakeholders; the importance of clarifying what specific behaviors describe "quality" in that context; and the value of revising, testing, and revising again. As well, it is important to note that country adaptations can also lead to deletions of constructs that could have great significance for child development, such as the decision to delete items indexing discipline and gender equity.

The emergence of a reliable factor structure indicates that MELE-A, when used in this country, satisfies one criterion for validity. As noted above, factor analyses of the CLASS and versions of the ERS have not found support for the standard three-factor structure. This indicates that the assumptions required for using the standard CLASS and ERS scoring procedures (i.e., using raw scores) are violated. Accordingly, the associations between classroom quality scores and child outcomes that have been presented in prior research may be biased due to inaccuracy (poor validity) and imprecision (poor reliability) in the classroom quality scores. Our results help establish a factor structure, which in turn should attenuate measurement error when reporting on associations with child outcomes. While we outline below several reasons why we may not have found significant associations with child development, including the possibility of measurement error, the factor structure provides some indication that MELE-A was functioning appropriately as a measure of ECE quality – which now should be replicated in other studies.

MELE-A scores demonstrated some validity evidence based on relations to other variables. Teachers' education levels were associated with all constructs of quality, and health/safety and materials/activities scores were higher on average for teachers who were less stressed. After accounting for a wide range of child, family and school characteristics, children who were in better organized classrooms had significantly higher scores on average on measures of child development using the MODEL direct assessment tool, although this association was small and was the only association uncovered among nine possible associations. In contrast, the teacher/child interaction and health/safety factors were not uniquely associated with any of the variables indexing child development and learning. The failure to find unique associations with either health/safety or teacher/child interaction was unexpected, especially given the critical importance of teacher/child interaction for child learning (e.g., Rogoff, 1990; Vygotsky, 1978).

Although the associations between quality and child development reported here are not markedly different than previous research (e.g., Perlman et al., 2016; Brunsek et al., 2017; Wolf et al., 2018), it is still notable that the quality factors explained a small amount of variance in children's learning as measured by direct assessment, and no significant variance in children's social and emotional learning. Our findings contrast with many other studies that have shown small to moderate associations of ECE on children's learning in LMIC (e.g., Rao et al., 2017). Scores on quality factors also showed no association with teachers' training; how-

ever, this study did not examine the content, dosage or intensity of that training, which may help explain why the associations were not found. In sum, this complex pattern of results is somewhat consistent with findings in United States (e.g., Keys et al., 2013), and forecasts the mixed results that may emerge when studying associations between aspects of quality and child development across large samples in LMIC.

There were several limitations to this study. Of central importance to this study are the possible limitations of the MELE-A. The decisions to delete or modify items to make them easier for observers may have obscured important variation in classroom settings or led to omissions of important constructs. The MELE-A may not have provided enough detail on variations of teacher/child interaction in this setting, leading to attenuated findings that underplay the significance of teachers. At the same time, it is also possible that other dimensions of ECE quality, such as children's engagement with materials, better differentiates quality in this setting than teacher/child interaction. Our study was not able to fully investigate whether MELE-A meaningfully differentiated between high- and low-quality ECE and on which dimensions the scale was most sensitive, another key marker of valid ECE quality measures (Mashburn, 2017). Issues with quality measurement more broadly, such as whether an accurate observation can be conducted in a short period of time and whether the constructs that are easily observed during an observation are the most critical for children's learning (Mashburn, 2017), also may have affected MELE-A's ability to accurately index classroom quality.

The data from MELE-A might have also failed to detect associations between quality and child development because the scale was not sensitive enough to identify thresholds of quality. Little is known about thresholds of quality necessary to support children's development in LMIC, which is an important gap given the resource constraints that may prohibit investments in ECE in many countries. Exposure to any formal learning environment may support child development in resource-constrained environments, leading to larger effect sizes than in high-income countries over time (Burchinal et al., 2011; Rao et al., 2012), or conversely, low-quality ECE could make little difference in children's learning (Sylva, Melhuish, Sammons, Siraj-Blatchford, & Taggart, 2011) or even inhibit healthy development.

There are other important questions on MELE-A's validity evidence that were not addressed by this study, for example, evidence of predictive validity or whether children who attend higher-quality classrooms as defined by the MELE-A perform better over time as reported by McCoy and Wolf (2018); whether the latent constructs identified here are consistent across samples; and the extent to which MELE-A is sensitive to intervention effects. Challenges to the implementation included the inability to thoroughly test and retest observer reliability and the failure to sample from all schools due to limitations in government listings. While statistical controls were used to account for family and classroom characteristics, it is possible that key influences on child development and learning were not captured in these models, leading to underestimation of the associations between quality and child development.

Our results suggest several avenues for future work in this area. A key question is how best to define and measure ECE quality and child development and learning in ways that are contextually relevant (Dahlberg et al., 1999), while remaining true to scientific findings on the importance of stimulating environments to support children's development. Government standards, used in this study to index context alignment, are just one view on what "quality" means to ECE stakeholders in any given country, and many important aspects of quality may not have been included in the MELE-A, either because the adaptation process was not expansive and in-depth enough, or because more research should be done in each country to generate definitions of quality that are local in nature, *before* quality tools are designed. Recommendations for future scale development include placing more emphasis on capturing teachers', parents' and other stakeholders' views of quality, and ensuring that scales are designed to capture nuances in quality that may differentiate between high- and low-quality settings even when resources are limited.

A second area for future work is to examine how the process of tool development – engaging stakeholders to address content of tools and ensure feasibility and alignment – affects *use of data* by stakeholders to improve policies and programs. This is a critical direction for future research, as large-scale measurement of quality is justified based on its potential contribution to ECE quality improvement. Beyond questions of psychometrics, data from MELE-A describe classrooms in this country and reveal several areas for improvement. For example, while teachers covered some topics such as mathematics frequently, other critical aspects of learning for young children, such as book reading, were infrequently observed. Most teachers were not using play-based learning strategies. Although health and safety did not emerge as an independent predictor of children's learning and development, the means on the items indicated that most children were not in facilities with running water or toilets. A next step is to examine if and how these findings encouraged changes in policy and practice in ECE.

To inform tool development and improve the evidence base, longitudinal and intervention studies are needed to examine impacts of ECE in samples of "typical" ECE settings in low-resource countries, characterized by little teacher training, inconsistent access to materials, and inadequate water and sanitation. The overall quality of evidence used to evaluate the associations between quality ECE and child development is not strong in LMIC (Rao et al., 2017), and this gap must be addressed to refine ECE programs, encourage routine government monitoring and support, and improve teacher training.

In sum, data and measurement can generate evidence to guide policymakers, teacher trainers, and other stakeholders on investment of limited resources in ECE. The tools available for low- and middle-income contexts are limited, and technical assistance for data collection and monitoring has been noted as a gap in the global infrastructure (Aboud & Proulx, 2019). The MELE represents a novel approach to quality measurement, by taking a "core" of items to adapt to better match local context. When viewed in the context of the mixed associations and variable factor structures from studies of more resource-intensive scales developed in high-income countries, the inconsistent associations do not seem unexpected, but raise several questions on MELE-A's potential usefulness to inform policy and practice, and more broadly, the importance of carefully examining the validity of ECE quality measurement of ECE in LMIC. Adaption of ECE quality measures is commonplace and likely inevitable to ensure both feasibility and appropriate test content, and more work should systematically document adaptation processes and resulting psychometric properties of measures. Such research is needed to minimize measurement error and effectively guide policy and practice at a critical time for ECE investments in many countries.

7. Funding

Support for this study was provided by the Children's Investment Fund Foundation and the Porticus Foundation.

We appreciate the extremely helpful comments offered by Dr. Rachel Gordon, Editor of this special edition, Dr. Kim Boller, and three anonymous reviewers.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.ecresq.2020.06.001.

References

Aboud, F. E., & Hossain, K. (2011). The impact of preprimary school on primary school achievement in Bangladesh. *Early Childhood Research Quarterly*, *26*(2), 237–246. http://dx.doi.org/10.1016/j.ecresq.2010.07.001

Aboud, F. E., & Proulx, K. (2019). *Strengthening early childhood care and education* Retrieved from:. Washington, DC: Global Partnership for Education. https://www.globalpartnership.org/sites/default/files/2019-007-17-kix-ecce-final-english.pdf

Administration for Children & Families. (2015). *QRS resource guide* Retrieved from:. https://qrisguide.acf.hhs.gov/sites/default/files/QRIS_Resource_Guide_2015.pdf

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Asparouhov, T., & Muthén, B. (2010). *Weighted least squares estimation with missing data* Retrieved from. Mplus webnote. https://www.statmodel.com/download/GstrucMissingRevision.pdf

Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin*, *107*, 238–246. http://dx.doi.org/10.1037//0033-2909.107.2.238

Bidwell, K., & Watine, L. (2014). *Exploring early education programs in peri-urban settings in Africa*. New Haven, CT: Innovations for Poverty Action.

Bietenbeck, J., Ericsson, S., & Wamalwa, F. M. (2019). Preschool attendance, schooling, and cognitive skills in East Africa. *Economics of Education Review*, *73*, Article 101909 http://dx.doi.org/10.1016/j.econedurev.2019.101909

Bihler, L. M., Agache, A., Kohl, K., Willard, J. A., & Leyendecker, B. (2018). Factor analysis of the Classroom Assessment Scoring System replicates the three domain structure and reveals no support for the bifactor model in German preschools. *Frontiers in Psychology*, *9* http://dx.doi.org/10.3389/fpsyg.2018.01232

Bornstein, M. H., Britto, P. R., Nonoyama-Tarumi, Y., Ota, Y., Petrovic, O., & Putnick, D. L. (2012). Child development in developing countries: Introduction and methods. *Child Development*, *83*(1), 16–31. http://dx.doi.org/10.1111/j.1467-8624.2011.01671.x

Brinkman, S. A., Hasan, A., Jung, H., Kinnell, A., Nakajima, N., & Pradhan, M. (2016). *The role of preschool quality in promoting child development: Evidence from rural Indonesia*. Washington, DC: The World Bank. http://dx.doi.org/10.1596/1813-9450-7529

Britto, P. R., Lye, S. J., Proulx, K., Yousafzai, A. K., Matthews, S. G., Vaivada, T., . . . & MacMillan, H. (2017). Nurturing care: Promoting early childhood development. *The Lancet*, *389*(10064), 91–102. http://dx.doi.org/10.1016/s0140-6736(16)31390-3

Britto, P. R., Yoshikawa, H., & Boller, K. (2011). Quality of early childhood development programs in global contexts: Rationale for investment, conceptual framework and implications for equity. Social policy report. *Society for Research in Child Development*, *25*(2) http://dx.doi.org/10.1002/j.2379-3988.2011.tb00067.x

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Browne, M. W., & Cudeck, R. (1993). Alternate ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage Publications.

Brunsek, A., Perlman, M., Falenchuk, O., McMullen, E., Fletcher, B., & Shah, P. S. (2017). The relationship between the Early Childhood Environment Rating Scale and its revised form and child outcomes: A systematic review and meta-analysis. *PloS One*.

Burchinal, M. (2018). Measuring early care and education quality. *Child Development Perspectives*, *12*(1), 3–9. http://dx.doi.org/10.1111/cdep.12260

Burchinal, M., Howes, C., Pianta, R., Bryant, D., Early, D., Clifford, R., . . . & Barbarin, O. (2008). Predicting child outcomes at the end of kindergarten from the quality of pre-kindergarten teacher–child interactions and instruction. *Applied Development Science*, *12*(3), 140–153. http://dx.doi.org/10.1080/10888690802199418

Burchinal, M., Kainz, K., & Cai, Y. (2011). How well do our measures of quality predict child outcomes? A meta-analysis and coordinated analysis of data from large-scale studies of early childhood settings. In M. Zaslow, I. Martinez-Beck, K. Tout, & T. Halle (Eds.), *Quality measurement in early childhood settings* (pp. 11–32). Washington, DC: Brookes Publishing.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Crouch, L., & Merseth, K. A. (2017). Stumbling at the first step: Efficiency implications of poor performance in the foundational first five years. *Prospects*, *47*(3), 175–196. http://dx.doi.org/10.1007/s11125-017-9401-1

Dahlberg, G., Moss, P., & Pence, A. R. (1999). *Beyond quality in early childhood education and care: Postmodern perspectives*. Levittown, PA: Psychology Press. http://dx.doi.org/10.4324/9780203980583

Denny, J. H., Hallam, R., & Homer, K. (2012). A multi-instrument examination of preschool classroom quality and the relationship between program, classroom, and teacher characteristics. *Early Education and Development*, *23*(5), 678–696. http://dx.doi.org/10.1080/10409289.2011.588041

Early, D. M., Maxwell, K. L., Burchinal, M., Alva, S., Bender, R. H., Bryant, D., . . . & Henry, G. T. (2007). Teachers' education, classroom quality, and young children's academic skills: Results from seven studies of preschool programs. *Child Development*, *78*(2), 558–580. http://dx.doi.org/10.1111/j.1467-8624.2007.01014.x

Early, D. M., Sideris, J., Neitzel, J., LaForett, D. R., & Nehler, C. G. (2018). Factor structure and validity of the early childhood environment rating scale—Third edition (ECERS-3). *Early Childhood Research Quarterly*, *44*, 242–256. http://dx.doi.org/10.1016/j.ecresq.2018.04.009

Garcia, M. H., Pence, A., & Evans, J. (Eds.). (2008). *Africa's future, Africa's challenge: Early childhood care and development in Sub-Saharan Africa*. The World Bank. http://dx.doi.org/10.1596/978-0-8213-6886-2

Garvis, S., Sheridan, S., Williams, P., & Mellgren, E. (2018). Cultural considerations of ECERS-3 in Sweden: A reflection on adaption. *Early Child Development and Care*, *188*(5), 584–593.

Gong, X., Xu, D., & Han, W. J. (2016). The effects of preschool attendance on adolescent outcomes in rural China. *Early Childhood Research Quarterly*, *37*, 140–152. http://dx.doi.org/10.1016/j.ecresq.2016.06.003

Gordon, R. A., Fujimoto, K., Kaestner, R., Korenman, S., & Abner, K. (2013). An assessment of the validity of the ECERS-R with implications for measures of child care quality and relations to child development. *Developmental Psychology*, *49*(1), 146. http://dx.doi.org/10.1037/a0027899

Hamre, B. K. (2014). Teachers' daily interactions with children: An essential ingredient in effective early childhood programs. *Child Development Perspectives*, *8*(4), 223–230. http://dx.doi.org/10.1111/cdep.12090

Hamre, B., Hatfield, B., Pianta, R., & Jamil, F. (2014). Evidence for general and domain-specific elements of teacher–child interactions: Associations with preschool children's development. *Child Development*, *85*(3), 1257–1274. http://dx.doi.org/10.1111/cdev.12184

Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A. J., Jones, S. M., . . . & Brackett, M. A. (2013). Teaching through interactions: Testing a developmental framework of teacher effectiveness in over 4,000 classrooms. *The Elementary School Journal*, *113*(4), 461–487. http://dx.doi.org/10.1086/669616

Hardin, B. J., Bergen, D., & Hung, H. F. (2013). Investigating the psychometric properties of the ACEI global guidelines assessment (GGA) in four countries. *Early Childhood Education Journal*, *41*(2), 91–101.

Harms, T., Clifford, R. M., & Cryer, D. (1980). *Early childhood environment rating scale*. New York: Teachers College Press.

Harms, T., Clifford, R. M., & Cryer, D. (1998). *Early childhood environment rating scale—Revised*. New York: Teachers College Press.

Harms, T., Clifford, R., & Cryer, D. (2015). *Early childhood environment rating scale* (3rd ed.). New York, NY: Teachers College Press.

Hu, B. Y. (2015). Comparing cultural differences in two quality measures in Chinese kindergartens: The Early Childhood Environment Rating Scale-Revised and the kindergarten quality rating system. *Compare: A Journal of Comparative Education*, *45*(1), 94–117.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55. http://dx.doi.org/10.1080/10705519909540118

Hu, B. Y., Fan, X., Gu, C., & Yang, N. (2016). Applicability of the classroom assessment scoring system in Chinese preschools based on psychometric evidence. *Early Education and Development*, *27*(5), 714–734. http://dx.doi.org/10.1080/10409289.2016.1113069

Jackson, J., Ahmed, S. K., Carslake, T., & Lietz, P. (2019). *Improving young children's learning in economically developing countries: What works, why, and where? Scoping review*. Camberwell, Australia: Australian Council for Educational Research. https://research.acer.edu.au/cgi/viewcontent.cgi?article=1038&context=monitoring_learning

Jamarillo, A., & Mingat, A. (2008). Early childhood care and education in sub-saharan Africa: what would it take to meet the millennium development goals? In M. Garcia, A. Pence, & J. Evans (Eds.), *Africa's Future, Africa's Challenge: Early childhood care and education in sub-Saharan Africa.*. Washington, DC: The World Bank.

Kaul, V., Chaudhary, A., & Sharma, S. (2014). *Indian early childhood education (IECEI) impact study – 1, quality and diversity in early childhood education—A view from Andhra Pradesh, Assam and Rajasthan. Centre for early childhood education and development*. New Delhi: Ambedkar University Delhi.

Keys, T. D., Farkas, G., Burchinal, M. R., Duncan, G. J., Vandell, D. L., Li, W., . . . & Howes, C. (2013). Preschool center quality and school readiness: Quality effects and variation by demographic and child characteristics. *Child Development*, *84*(4), 1171–1190. http://dx.doi.org/10.1111/cdev.12048

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford Press.

Kotzé, J. (2015). The readiness of the South African education system for a pre-grade R year. *South African Journal of Childhood Education*, *5*(2), 1–27. http://dx.doi.org/10.4102/sajce.v9i1.597

Leyva, D., Weiland, C., Barata, M., Yoshikawa, H., Snow, C., Treviño, E., . . . & Rolla, A. (2015). Teacher–child interactions in Chile and their associations with prekindergarten outcomes. *Child Development*, *86*(3), 781–799. http://dx.doi.org/10.1111/cdev.12342

Li, H., Liu, J., & Hunter, C. V. (2019). A meta-analysis of the factor structure of the classroom assessment scoring system (CLASS). *The Journal of Experimental*

Education, http://dx.doi.org/10.1080/00220973.2018.1551184
Malmberg, L. E., Mwaura, P., & Sylva, K. (2011). Effects of a preschool intervention on cognitive development among East-African preschool children: A flexibly time-coded growth model. *Early Childhood Research Quarterly*, *26*(1), 124–133. http://dx.doi.org/10.1016/j.ecresq.2010.04.003
Mariano, M., Caetano, S. C., Ribeiro da Silva, A., Surkan, P. J., Martins, S. S., & Cogo-Moreira, H. (2019). Psychometric properties of the ECERS-R among an epidemiological sample of preschools. *Early Education and Development*, *30*(4), 511–521.
Martinez, S., Naudeau, S., & Pereira, V. (2012). *The promise of preschool in Africa: A randomized impact evaluation of early childhood development in rural Mozambique*. Washington D.C: The World Bank. http://siteresources.worldbank.org/INTAFRICA/Resources/The_Promise_of_Preschool_in_Africa_ECD_REPORT.pdf
Mashburn, A. J. (2017). Evaluating the validity of classroom observations in the head start designation renewal system. *Educational Psychologist*, *52*(1), 38–49.
McCoy, D. C., & Wolf, S. (2018). Changes in classroom quality predict Ghanaian preschoolers' gains in academic and social-emotional skills. *Developmental Psychology*, *54*(8), 1582. http://dx.doi.org/10.1037/dev0000546
Montes, G., Reynolds Weber, M., Infurna, C., Van Wagner, G., Zimmer, A., & Hightower, A. D. (2018). Factor structure of the ECERS-3 in an urban setting: An independent, brief report. *European Early Childhood Education Research Journal*, *26*(6), 972–984.
Montie, J. E., Xiang, Z., & Schweinhart, L. J. (2006). Preschool experience in 10 countries: Cognitive and language performance at age 7. *Early Childhood Research Quarterly*, *21*(3), 313–331. http://dx.doi.org/10.1016/j.ecresq.2006.07.007
Moore, A. C., Akhter, S., & Aboud, F. E. (2008). Evaluating an improved quality preschool program in rural Bangladesh. *International Journal of Educational Development*, *28*(2), 118–131. http://dx.doi.org/10.1016/j.ijedudev.2007.05.003
Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide (Version 7)*. Los Angeles, CA: Author.
Mwaura, P. A., Sylva, K., & Malmberg, L. E. (2008). Evaluating the Madrasa preschool programme in East Africa: a quasi-experimental study. *International Journal of Early Years Education*, *16*(3), 237–255.
Myers, R. G. (2004). *In search of quality in programs of early childhood care and education. Background paper for education for all, global monitoring report 2005*. Paris: UNESCO. www.unesco.org/education/gmr_download/references_2005.pdfS
Neuman, M. J., & Okeng'o, L. (2019). Early childhood policies in low-and middle-income countries. *Early Years Journal*, *39*(3), 233–238.
Nonoyama-Tarumi, Y., Loaiza, E., & Engle, P. (2009). Inequalities in attendance in organized early learning programmes in developing societies: Findings from household surveys. *Compare: A Journal of Comparative and International Education*, *39*(3), 385–409. http://dx.doi.org/10.1186/1471-2458-13-1049
OECD. (2015). *Starting strong IV: Monitoring quality in early childhood education and care*. Paris: Starting Strong, OECD Publishing. http://dx.doi.org/10.1787/9789264233515-en
Pakarinen, E., Lerkkanen, M. K., Poikkeus, A. M., Kiuru, N., Siekkinen, M., Rasku-Puttonen, H., . . . & Nurmi, J. E. (2010). A validation of the classroom assessment scoring system in Finnish kindergartens. *Early Education and Development*, *21*(1), 95–124.
Perlman, M., Falenchuk, O., Fletcher, B., McMullen, E., Beyene, J., & Shah, P. S. (2016). A systematic review and meta-analysis of a measure of staff/child interaction quality (the classroom assessment scoring system) in early childhood education and care settings and child outcomes. *PLoS One*, *11*(12), Article e0167660 http://dx.doi.org/10.1371/journal.pone.0167660
Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system™: Manual Pre-K*. Baltimore, MD: Paul H Brookes Publishing.
Ponguta, L. A., Maldonado-Carreño, C., Kagan, S. L., Yoshikawa, H., Nieto, A. M., Aragón, C. A., . . . & Guerrero, P. A. (2019). Adaptation and application of the Measuring Early Learning Quality and Outcomes (MELQO) Framework to early childhood education settings in Colombia. *Zeitschrift für Psychologie*, *227*(2), 105–112. http://dx.doi.org/10.1027/2151-2604/a000361
Proulx, K., & Aboud, F. (2019). Disaster risk reduction in early childhood education: Effects on preschool quality and child outcomes. *International Journal of Educational Development*, *66*, 1–7.
Rao, N., Sun, J., Pearson, V., Pearson, E., Liu, H., Constas, M. A., . . . & Engle, P. L. (2012). Is something better than nothing? An evaluation of early childhood programs in Cambodia. *Child Development*, *83*(3), 864–876. http://dx.doi.org/10.1111/j.1467-8624.2012.01746.x
Raikes, A., Davis, D., & Burton, A. (2019). Early Childhood Care and Education in the Era of Sustainable Development: Balancing Local and Global Priorities. In L. Suter, E. Smith, & B. Denman (Eds.), *The SAGE Handbook of Comparative Studies in Education*. New York: Sage.
Raikes, A., Koziol, N., Janus, M., Platas, L., Weatherholt, T., Smeby, A., & Sayre, R. (2019). Examination of school readiness constructs in Tanzania: Psychometric evaluation of the MELQO scales. *Journal of Applied Developmental Psychology*, *62*, 122–134.
Raikes, A., Sayre, R., Davis, D., Anderson, K., Hyson, M., Seminario, E., & Burton, A. (2019). The Measuring Early Learning Quality & Outcomes initiative: purpose, process and results. *Early Years*, *39*(4), 1–16. http://dx.doi.org/10.1080/09575146.2019.1669142
Raikes, A., Yoshikawa, H., Britto, P., & Iruka, I. (2017). Children, youth and developmental science in the Sustainable Development Goals. Social Policy Report. *Society for Research in Child Development*, *30*(2) http://dx.doi.org/10.1002/j.2379-3988.2017.tb00088.x
Rao, N., Sun, J., Chen, E. E., & Ip, P. (2017). Effectiveness of early childhood interventions in promoting cognitive development in developing countries: A systematic review and meta-analysis. *Hong Kong Journal of Paediatrics*, *22*(1), 14–25.
Rogoff, B. (1990). *Apprenticeship in thinking: Cognitive development in social context*. Oxford university press.
Rossiter, J., Hagos, B., Rose, P., Teferra, T., & Woldehanna, T. (2018). *Early learning in Ethiopia: Equitable access and learning. System diagnostic report for world bank early learning program*. https://www.educ.cam.ac.uk/centres/real/downloads/ELP%20System%20Diagnostic%20Final_Nov%202018_updated.pdf
RTI. (2018). *Measuring early learning and quality outcomes (Final report)*. Mainland Tanzania: The World Bank. https://www.brookings.edu/wp-content/uploads/2017/06/melqo-measuring-early-learning-quality-outcomes-in-tanzania_2016oct.pdf
Sabanathan, S., Wills, B., & Gladstone, M. (2015). Child development assessment tools in low-income and middle-income countries: How can we use them more appropriately? *Archives of Disease in Childhood*, *100*(5), 482–488. http://dx.doi.org/10.1136/archdischild-2014-308114
Sandell, E. J., Hardin, B. J., & Wortham, S. C. (2010). Using ACEI's global guidelines assessment for improving early education. *Childhood Education*, *86*(3), 157–160.
Sandilos, L. E., & DiPerna, J. C. (2014). A review of empirical evidence and practical considerations for early childhood classroom observation scales. *NHSA Dialog*, *17*(2) https://www.crcpress.com/rsc/downloads/Early_Years_Making_it_Count.pdf
Seidman, E., Raza, M., Kim, S., & McCoy, J. M. (2014). *Teacher instructional practices and processes system–TIPPS: Manual and scoring system*. New York: New York University.
Shallwani, S., Abubakar, A., & Kachama, M. (2018). The quality of learning and care at community-based early childhood development centers in Malawi. *Global Education Review*, *5*(2), 28–46.
Sheridan, S., Giota, J., Han, Y. M., & Kwon, J. Y. (2009). A cross-cultural study of preschool quality in South Korea and Sweden: ECERS evaluations. *Early Childhood Research Quarterly*, *24*(2), 142–156. http://dx.doi.org/10.1007/2288-6729-3-1-1
Singh, R., & Mukherjee, P. (2018). Effect of preschool education on cognitive achievement and subjective wellbeing at age 12: Evidence from India. *Compare: A Journal of Comparative and International Education*, 1–19. https://www.younglives.org.uk/sites/www.younglives.org.uk/files/YL-WP175-Singh.pdf
Sylva, K., Siraj-Blatchford, I., & Taggart, B. (2003). *Assessing quality in the early years: Early childhood environment rating scale: Extension (ECERS-E), four curricular subscales*. Trentham Books.
Sylva, K., Melhuish, E., Sammons, P., Siraj-Blatchford, I., & Taggart, B. (2011). Pre-school quality and educational outcomes at age 11: Low quality has little benefit. *Journal of Early Childhood Research*, *9*(2), 109–124. http://dx.doi.org/10.1177/1476718X10387900
Sylva, K., Siraj-Blatchford, I., Taggart, B., Sammons, P., Melhuish, E., Elliot, K., . . . & Totsika, V. (2006). Capturing quality in early childhood through environmental rating scales. *Early Childhood Research Quarterly*, *21*(1), 76–92. http://dx.doi.org/10.1016/j.ecresq.2006.01.003
Thornburg, K. R., Mauzy, D., Mayfield, W. A., Hawks, J. S., Sparks, A., Mumford, J. A., . . . & Fuger, K. L. (2011). Data-driven decision making in preparation for large-scale quality rating system implementation. In M. Zaslow, I. Martinez-Beck, K. Tout, & T. Halle (Eds.), *Quality measurement in early childhood settings* (pp. 363–388). Washington, DC: Brookes Publishing.
UNESCO, UNICEF, Brookings Institution, & The World Bank. (2017). *Overview MELQO: Measuring early learning quality outcomes*. Paris: UNESCO.
UNESCO Institute of Statistics. (2019). *Country statistics* Retrieved from:. http://uis.unesco.org/
United Nations Development Program. (2019). *Human development index country profiles* Retrieved from. http://hdr.undp.org/en/countries
Urban, M. (2019). The shape of things to come and what to do about Tom and Mia: Interrogating the OECD's International Early learning and Child Well-Being Study from an anti-colonialist perspective. *Policy Futures in Education*, *17*(1), 87–101. http://dx.doi.org/10.1177/1478210318819177
Vermeer, H. J., van IJzendoorn, M. H., Cárcamo, R. A., & Harrison, L. J. (2016). Quality of child care using the environment rating scales: A meta-analysis of international studies. *International Journal of Early Childhood*, *48*(1), 33–60. http://dx.doi.org/10.1007/s13158-015-0154-9
Vygotsky, L. (1978). Interaction between learning and development. *Readings on the Development of Children*, *23*(3), 34–41.
Wolf, S., Raza, M., Kim, S., Aber, J. L., Behrman, J., & Seidman, E. (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. http://dx.doi.org/10.1016/j.ecresq.2018.05.003
Wuermli, A. J., Tubbs, C. C., Petersen, A. C., & Aber, J. L. (2015). Children and youth in low- and middle-income countries: Toward an integrated developmental and intervention science. *Child Development Perspectives*, *9*(1), 61–66. http://dx.doi.org/10.1111/cdep.12108
Yoshikawa, H., Wuermli, A. J., Raikes, A., Kim, S., & Kabay, S. B. (2018). Toward high-quality early childhood development programs and policies at national scale: Directions for research in global contexts. *Social Policy Report*, *31*(1), 1–36.