IDB WORKING PAPER SERIES Nº IDB-WP-742

The Psychometric Properties of Child Care Quality Measures in Ecuador: Lessons for Monitoring Process Variables at Scale

Florencia Lopez Boo
Marta Dormal
Ann Weber

Inter-American Development Bank
Social Protection and Health Division

October 2016

IDB
Inter-American
Development Bank

The Psychometric Properties of Child Care Quality Measures in Ecuador: Lessons for Monitoring Process Variables at Scale

Florencia Lopez Boo
Marta Dormal
Ann Weber

October 2016

Cataloging-in-Publication data provided by the
Inter-American Development Bank
Felipe Herrera Library
Lopez Boo, Florencia.
The psychometric properties of child care quality measures in Ecuador: lessons for monitoring process variables at scale / Florencia Lopez Boo, Marta Dormal, Ann Weber.
p. cm. — (IDB Working Paper Series ; 742)
Includes bibliographic references.
1. Child care-Ecuador-Evaluation. 2. Child development-Ecuador-Evaluation. 3. Day care centers-Ecuador-Evaluation. I. Dormal, Marta. II. Weber, Ann. III. Inter-American Development Bank. Social Protection and Health Division. III. Title. IV. Series.
IDB-WP-742

scl-sph@iadb.org

www.iadb.org/SocialProtection

The Psychometric Properties of Child Care Quality Measures in Ecuador: Lessons for Monitoring Process Variables at Scale

Florencia Lopez Boo[*], Marta Dormal[*] and Ann Weber[†]

Abstract

This paper analyzes four instruments that are widely used to measure the quality of centers serving children ages 0 to 36 months - the CLASS, the ITERS-R, the CC-IT-HOME, and the MITRCC – and that were administered to a sample of 404 child care centers in Ecuador. We first assess the psychometric properties of these instruments in their first application in Ecuador. Specifically, we examine their internal consistency, test the underlying subscale structure by means of confirmatory factor analysis (CFA), verify construct validity by testing associations with quality-related factors (e.g., child-caregiver ratio), and check concurrent validity of the instruments' total scores. We then explore how we can use the data from these instruments to inform the development of a simple, less costly checklist that programs in Latin America and the Caribbean (LAC) can employ to monitor service quality. To that end, we look at the correlation of the separate subscales from the simpler measures (ITERS-R, CC-IT-HOME, and MITRCC) with each of the dimensions of the CLASS. In addition, we map the subscales of all of the instruments to a set of child-caregiver interaction constructs identified in the literature as core for the healthy development of children, and used CFA to test how well the instrument subscales loaded onto each latent construct separately. Overall, the instruments showed excellent consistency. The first CFA also demonstrated that the data were a good fit to the published structure of each instrument. Associations with quality-related factors exhibited the expected signs, and concurrent validity across instruments revealed low correlations between overall scores on CLASS and the other instruments. Moderate correlations were found between certain dimensions of CLASS and the subscales of other instruments that reflected *a priori* similar constructs (in particular the Listening and Talking, Interaction, and Personal Care Routines subscales of the ITERS-R). Finally, the second CFA revealed that dimensions of CLASS had the highest loadings to three of the theoretical constructs for child-caregiver interaction: Sensitivity/Responsiveness, Positive Regard/Warmth and Joint Attention. However, several of the subscales of the ITERS-R and CC-IT-HOME, as well as the MITRCC showed encouraging associations with theoretical constructs for process quality that were not as strongly captured by CLASS.

JEL codes: I10, I20, I30, J13, I38

Keywords: child development, child care centers, quality measures, daycare centers, correlations, validity, psychometric properties, CLASS, ITERS-R, CC-IT-HOME, MITRCC.

[*] Social Protection and Health Division, Inter-American Development Bank, Washington DC, USA.
[†] Department of Pediatrics, Stanford University, Stanford, California, USA.

We, the authors, would like to thank the team of specialists from the Inter-American Development Bank, consisting of María Caridad Araujo, Rafael Novella, Sara Schodt, and particularly Romina Tomé at Duke University. They all participated in one or more of the various stages of study design, data collection and data analysis. We would also like to thank the technical staff from INFA, MIES and MCDS in Ecuador for their unconditional support, and the team from Habitus Research for their excellent work throughout the data collection phase and on the coding of CLASS videos. We would also like to acknowledge Dr. Richard Clifford, Dr. Kathy Thornburg, Dr. Robert Bradley, Dr. Bridget Hamre, Dr. Jennifer LoCasale-Crouch, Dr. Tamara Halle, Dr. Sally Atkins-Burnett and Lorena Sernett for their feedback on the child-caregiver interaction constructs. All errors are our own.

1 Introduction

In recent years, several countries in Latin America and the Caribbean (LAC) have made a considerable investment in expanding child care coverage for young children (available in daycare centers, nursery schools and preschools); however, we know that the quality of these services offered in LAC is quite low (Berlinski and Schady, 2015; Araujo et al., 2015; Verdisco et al., 2010).[1] Countries in the region now face the challenge of ensuring the quality of the child care services they offer, especially since the children who attend these centers do so during a critical period in their development.

Several experimental and quasi-experimental studies, particularly those conducted in the US and other developed countries, have shown that the impact of child care centers on child development mostly depends on the quality of the centers, especially their process quality (López-Boo et al., 2016). Process variables,[2] which characterize the quality of child care routines and interactions between children and caregivers, become even more important for children under the age of 3 who are in the process of forming an attachment to a significant adult (Bowlby, 1969).[3] Furthermore, although young children require less structured curricular content, they demand more individualized attention than older children because they depend more heavily on the caregiver to initiate an interaction until they acquire full mobility (Howes et al., 1992). Researchers have found that children in child care environments characterized by high process quality are not only able to initiate and engage in higher-order learning with their peers but also to achieve higher scores on academic achievement tests at a later age (Campbell and Ramey, 1995). The positive effects of high-quality programs tend to persist over time and usually last into adulthood (Vandell et al., 2010; Greenberg, Domitrovich and Bumbarger, 2001; Hamre and Pianta, 2007; Shonkoff and Phillips, 2000; Peisner-Feinberg et al., 2001; Burchinal et al., 1997). Additionally, the literature reveals that process indicators measuring the quality of child care centers are much more consistently related to the overall quality of care and children's developmental outcomes than are structural indicators[4] (La Paro et al., 2004; Hamre and Pianta, 2007; Mashburn et al., 2008; Li et al., 2013).

Nonetheless, the same process variables that are vital to ensuring quality care for infants and toddlers represent one of the region's greatest shortcomings (Berlinski and Schady, 2015) and are also the most difficult variables to measure. These variables require expert observation, judgment and interpretation, which is why it is so complex, time-consuming and costly to measure them. This partly explains why child care services in the region, as part of

[1] A recent evaluation of the U.S. Early Head Start centers, which serve children under the age of 3 from low-income families, shows that these centers fall in the moderate to high-quality range on the Toddler CLASS instrument (Vogel et al., 2015). These scores are slightly higher than those found in other LAC countries, including Ecuador.

[2] Process variables describe the quality of interactions between the children and their caregivers and between the children themselves, as well as the activities they engage in while at the child care center.

[3] Within the 0-3 year old age group, however, the evidence as to whether process quality is more important for infants vs. toddlers (i.e., under 1 year olds vs. 2 to 3 years olds) in determining later outcomes is, to the authors' knowledge, nonexistent. According to Halle et al. (2011), there are in fact a very limited number of measures that categorize items (and constructs) as being suitable for infants vs. toddlers. As such, they suggest that this type of analysis would require examining quality measures at the item level. On the other hand, they argue that even instruments that are able to make such a distinction at the item level may not have predictive validity for those relevant items.

[4] Structural variables generally identify the resources that facilitate those interactions: group size; the caregiver's education, experience and salary; infrastructure and safety; curriculum; and materials.

their ongoing monitoring and follow-up strategies, choose to capture structural quality variables, using checklist measures that primarily focus on easily quantifiable aspects of care (e.g., the number of books in the classroom), instead of process quality variables.

In an attempt to fill a void in the LAC literature, Araujo et al. (2015) present a detailed analysis of different aspects of service quality at *Centros Infantiles del Buen Vivir* (CIBVs) in Ecuador. In 2012, researchers administered a battery of instruments that had already been used in other countries in the region to measure the quality of child care services provided to children under 3 years of age. Their study included a brief analysis of the correlations between different instruments, revealing low correlations between overall scores on more specialized, complex instruments for measuring child care quality and scores on low-cost, less complex instruments. We build on their work by analyzing in more detail the performance of these instruments in the Ecuadorian context.

The objectives for this paper are two-fold. First, we aim to assess the psychometric properties of four well-established instruments to measure child care quality – the CLASS, the ITERS-R, the CC-IT-HOME, and the MITRCC - in their first application in Ecuador. For each instrument, we examine internal consistency, test the underlying subscale structure by means of confirmatory factor analysis (CFA), and verify construct validity by testing associations with quality-related factors, such as the child-caregiver ratio, location of the center in an urban vs. rural area, and cost to families for their child to attend. Also for this first objective, we check the concurrent validity of the instruments' total scores on the same CIBV sample.

For our second objective, we explore how we can use the data from the four instruments to inform the development of a simple, less costly checklist that programs in LAC can employ to monitor, supervise and track the quality of their service providers within the context of evidence-based quality improvement.[5] With this second goal in mind, we relied throughout the paper on a ranking of measures from least to most complex, in terms of both the administration of the instrument and its ability to capture process quality. Based on this ranking, we selected the CLASS as the reference standard for comparison because, to date, CLASS is the only tool that meets the following criteria: (i) it predicts the development of preschool-age children better than other instruments (Mashburn et al., 2008); (ii) it focuses exclusively on process elements (especially child-caregiver interactions), having been designed to address the limitations of other instruments focused on structural variables; (iii) it has been widely validated in several different contexts and in many countries, including LAC countries; and (iv) it requires that the trainer of the observers participate in formal training. Furthermore, in the context of this study, CLASS has the strictest administration protocols, both prior to and during fieldwork (in terms of inter-rater reliability standards, observer profiles, and group selection) as well as post-fieldwork (e.g., double coding of videos).[6]

To achieve our second objective, we look at the correlation of the separate subscales from the simpler measures (ITERS-R, CC-IT-HOME, and MITRCC) with each of the dimensions of the CLASS. In addition, we build a theoretical measurement model in which we map the

[5] See the quality rating and improvement systems (QRIS) used in the US for an understanding of quality improvement systems based on continuous measurement.
[6] See more details in Araujo et al. (2015).

subscales of all of the instruments to a set of latent constructs for child-caregiver interaction identified in the literature as core for the healthy development of children. We carefully examined the correlations between subscales and their *a priori* matched constructs, and used CFA to test the fit of the instrument subscales to the process quality model.

Apart from a study by Kane and Staiger (2012),[7] we found no other studies in LAC or other countries that have undertaken this type of exercise, precisely because it is very difficult to administer such a diverse set of measures to a single sample. For this reason, we believe that the validity exercise conducted in this paper, combined with other complementary results we provide, will inform the search for cost-effective process quality indicators for their frequent administration at scale.

This paper is organized into five parts. The second section describes the context of the CIBVs. The third presents the sample for analysis and the instruments used to measure child care quality, as well as describes the methodology used in this study. The fourth part presents the main results of the study, and the fifth section closes with a discussion of the results and the conclusion.

2 Context

The CIBVs are the primary providers of public child care services in Ecuador. According to CIBV administrative data, in 2011 these centers served some 140,000 children at approximately 3,800 centers throughout the country (Araujo, López-Boo and Puyana, 2013). The service mainly operates under third-party agreements with local governments, community organizations, foundations, churches, etc., which receive a transfer of public resources to cover the service's operating costs. Some of these entities supplement public funds with their own resources. Unlike most child care services in the region, at the time of data collection, this program functioned in both rural and urban areas. Although caregivers are officially required to have completed secondary school, in practice, compliance with this requirement is lax. Caregivers are hired by the organization, which acts as an operator, and they earn the minimum wage.

There are two main challenges facing these centers when it comes to providing high-quality care, and it is worth briefly highlighting them in order to understand the context of quality measurement in this paper. First, centers group children into classrooms with a very broad range of ages, and they also serve children over 3 years of age. During fieldwork, researchers encountered that only two of the 404 centers visited (0.5%) followed established program guidelines for all of their groups. At 93% of the centers, more than 50% of the groups of children did not conform to the age range established by the guidelines. The fact that the children were not grouped by age causes the age composition of the classrooms studied to be more heterogeneous than planned. In a child care setting like that of the CIBVs, where the staff members in charge of the children are not professionals, this mix of ages further complicates the task of child care. The second challenge relates to the quality of

[7] Our study, however, is different, since our dataset contains no information on child outcomes or feedback from children. In Kane and Staiger (2012), Table 10 shows relatively high correlations between five instruments (CLASS, FFT, MQI, PLATO, UTOP).

staff (coordinators and caregivers) working at the CIBVs, in terms of their knowledge and specific relevant skills. At the time of data collection for this study, the hiring of professionals as coordinators had been implemented at almost all of the centers studied. Nevertheless, and despite the fact that the CIBVs largely employ coordinators with post-secondary education, service quality ranks consistently low. Caregivers are only required to have completed secondary school, a requirement that, on average, is not met and that is reflected in the scores of the quality measures presented in section 4. Further details on the characteristics of the centers, the population and the sample are reported in Table A1 of the annex.

These two issues have certainly influenced the scores obtained for the quality measures, which overall reflect a low quality context (i.e., most of the scores are in the bottom 30% of the range of possible scores). Nonetheless, there exists substantial variation in scores within this lower range of the distribution, which we are able to exploit for our analysis. The results found here may differ in a context of higher quality than that of the centers in our sample.

3 Methods

Sample

The population from which the sample for this study was selected consisted of all child care centers in the administrative databases provided to the Inter-American Development Bank (IDB) by the Institute for Children and Families (INFA)[8] in May 2012. That year, these databases contained information on 3,575 centers, including data on the number of children enrolled and the number of community staff that worked at these centers.[9] The sample was stratified into two groups: centers with high child-caregiver ratios (1,779 centers had 9.2 children per adult or more – the median ratio in the administrative data) and centers with low child-caregiver ratios (1,776 centers had fewer than 9.2 children per adult).[10, 11] Based on this population, about 200 centers were randomly chosen from each of the two aforementioned groups, for a final sample of 404 centers.

During the fieldwork phase, each of the centers in the sample was visited for a full day by a pair of researchers responsible for collecting data on one group of children[12] and their caregiver(s), per center, and for administering the quality measures. All of the measures were administered in the same order and at the same time of day to ensure comparability across centers, as shown in Figure 1. In terms of the researchers' profile, each pair of interviewers that visited the CIBVs included an experienced interviewer with complete secondary education, who was responsible for filming the group, administering the MITRCC

[8] At the time of the study, the CIBVs came under the auspices of INFA, which was later incorporated into Ecuador's Ministry of Economic and Social Inclusion following the completion of data collection.
[9] These 3,575 centers represent about 92.4% of the Institute's CIBV population.
[10] The specialized literature identifies child-caregiver ratios as a key structural variable associated with the quality of child care.
[11] We replicate this stratification using child-caregiver ratios calculated on the basis of data collected at the CIBVs for this analysis. Table A2 of the annex presents the average scores on quality measures by type of center, and the p-value for the null hypothesis that the difference between the scores of the two groups is equal to zero. The results show that on some scales, the average score of centers with a low child-caregiver ratio (high-quality centers) is significantly higher than the score for centers with a high child-caregiver ratio (low-quality centers).
[12] The reference group was composed of children who were under the age of 36 months at the beginning of the school year. If there was more than one group of children in this age range, priority was given to the one in which all of the children fell within that range. See more details on sample selection and the fieldwork phase in Araujo et al. (2015).

and conducting interviews, and a researcher with post-secondary education in the field of child psychology or early childhood education, who was responsible for the administration of the CC-IT-HOME and ITERS-R (profiles 1 and 2, respectively, in Figure 1). CLASS administration was performed using classroom video footage of a four-hour day, from which four 20-minute segments were extracted.[13] The field research team (profile 1) was solely responsible for shooting the videos that were subsequently evaluated by a team of certified CLASS coders with the same profile as the profile 2 researchers.

[Figure 1. Protocol for instrument administration/coding in each center]

Instruments

Of the measures administered, this study focuses on the four internationally validated instruments that are characterized by providing a global measurement of child care quality:

1. The Classroom Assessment Scoring System for Toddlers (Pianta, La Paro and Hamre, 2008) or Toddler CLASS (CLASS from now on), which explores eight dimensions: Positive Climate, Negative Climate, Teacher Sensitivity, Regard for Child Perspectives, and Behavior Guidance are all grouped under the Emotional and Behavioral Support domain, while Facilitation of Learning and Development, Quality of Feedback, and Language Modeling are grouped under the Engaged Support for Learning domain;

2. The Infant/Toddler Environment Rating Scale – Revised Edition (ITERS-R) (Harms, Cryer and Clifford, 2006), herein referred to as ITERS-R, which is composed of seven subscales: Space and Furnishings, Personal Care Routines, Listening and Talking, Activities, Interaction, Program Structure, and Parents and Staff*;*

3. The Child Care Infant/Toddler Home Observation for Measurement of the Environment (CC-IT-HOME) (Bradley, Caldwell and Corwyn, 2003), herein referred to as CC-IT-HOME, which is organized into six subscales: Caregiver Responsivity, Acceptance, Organization, Learning Materials, Caregiver Involvement, and Variety of Stimulation; and

4. The Missouri Infant/Toddler Responsive Caregiving Checklist (MITRCC) (MU Center for Family Policy and Research, 2003), which contains 20 items organized into three groups of child development being promoted through quality interactions: social-emotional, physical and cognitive.

Since the ultimate goal of this analysis is to provide information for a tool of continuous quality monitoring of child care services, we present below the characteristics of the instruments that underlie their level of complexity, that is: (i) the type of variables measured (structural vs. process); (ii) the method of measuring process quality (observation vs. report); (iii) the instrument's training, scoring method, construct measured by the subscales/dimensions, and administration time; and (iv) the instrument's administration

[13] Segments were selected following the CLASS editing protocol. This protocol was not included in order to keep the document to a reasonable length; however, it is available to interested readers. Please contact the authors at florencial@iadb.org.

cost.[14] Based on the characteristics presented below, for the purposes of this analysis, we refer to CLASS as the complex measure and to the other instruments as simple measures. Table 1 presents the four instruments' main characteristics.

[Table 1. Main characteristics of the child care quality measures]

With regard to the type of variables measured, CLASS is the only one out of the four instruments that focuses exclusively on process variables (child-caregiver interactions, specifically), an important distinction given that, as mentioned in the introduction, process variables have a greater impact on child development. In contrast, the ITERS-R is composed of one structural subscale (Space and Furnishings) and six process subscales. It is important to note that only two subscales of the ITERS-R emphasize child-caregiver interactions and language (Interaction and Listening and Talking, respectively), with the remaining subscales focused primarily on materials, curriculum/schedule, and routines with parents and staff. The CC-IT-HOME has one dimension that focuses on learning materials, while the rest focus on processes, such as Caregiver Responsivity. Lastly, the 20 items of the MITRCC cover both process (e.g., the caregiver's recognition of and response to children's verbal and nonverbal cues) and structural aspects (e.g., items such as "Caregivers organize the classroom so that children have an opportunity to observe their surroundings from more than one level.").

While all of the instruments assess, at least in part, CIBV process quality, the mechanism for collecting data varied from center to center. Of the instruments presented, only CLASS and the MITRCC collect data exclusively through observation. The other instruments combine observation with information collected through reporting, in which a structured interview or survey is administered to a qualified informant (in this case, the caregiver in the classroom under study), using a list of questions about the dimensions of quality being evaluated.[15] This distinction is important because observation (as long as it is properly conducted) is the best method for describing the experiences and interactions of children at the center. Furthermore, this method usually produces data with a higher level of objectivity than that obtained through reporting, which may be subject to bias. For example, a caregiver may be unwilling to share information that could reveal that the attention she provides to the children in her care is less than optimal, or a person with a low level of education may be unable to accurately report on the frequency with which she carries out activities with the child. In fact, empirical evidence shows that direct observation instruments used to evaluate service quality predict child outcomes to a significantly greater degree than interviews or checklists (Zaslow et al., 2006). It is important to highlight that, although data collection for the MITRCC, CC-IT-HOME and ITERS-R was conducted through firsthand observation (i.e., direct observation performed during center visits), CLASS data collection was conducted by filming the experience of the children for later coding, which has even more advantages in terms of data quality.[16]

[14] A detailed description of these four instruments and their characteristics, costs, advantages and disadvantages is presented in López-Boo et al. (2016).
[15] The ITERS-R involves a minimum of three hours of observation and coding, plus 20 to 30 minutes of interviews. In the case of the CC-IT-HOME, 14 items (out of a total of 42) are completed during a survey given to the caregiver.
[16] One of these advantages is that the video coder is able to focus on what is happening with the child, blocking out what is going on around her (e.g., other groups of children who are not part of the sample or the activities of support staff).

Table 1 also shows how the complexity of administration varies with each instrument. Although the use of observation as a data collection method usually produces better quality data (see section 4), it also requires more time and training. The observer must be trained in the use of the instrument; capable of accurately capturing the interactions, routines, bonds, stimuli and activities that must be reported, without becoming distracted by inconsequential details; and able to objectively document, code and/or score each dimension. Of the four instruments analyzed, only CLASS requires formal training. Specifically, each coder is required to be certified on the instrument (valid for one year) by participating in a two-day training and passing a reliability test. The completion of formal training is not required to administer the other measures; however, user guides generally recommend that evaluators participate in training and achieve reliability in the coding process, especially in the case of the ITERS-R.[17]

Although the educational requirements for observers are similar (post-secondary education is usually required), training requirements generally reflect the complexity of the scoring method. In the case of CLASS, for example, each dimension is assigned a score ranging from 1 to 7. The assigned score depends on a series of assessments made by the observer, guided by the manual and knowledge gained during training so as to maintain objectivity and accuracy. The ITERS-R also has a relatively complex scoring system, which explains why participation in a training course led by a trainer experienced in the use of this instrument is usually recommended. In fact, it is designed to be administered gradually, meaning that once a classroom fails to comply with a subset of items within a given subscale, it receives the score corresponding to the highest stop point. Each of the 39 items that compose the instrument's seven dimensions is scored on a 7-point scale. In contrast, the CC-IT-HOME and MITRCC use a checklist format, with a binary (yes/no) response for each item. On the CC-IT-HOME, each subscale receives a score based on the sum of the individual items. These scores are then combined for a maximum possible total score of 42. On the MITRCC, the number of yes responses (between 0 and 20) is divided by two in order to arrive at an overall score that falls between 0 and 10.

Although they are more difficult to score, instruments that use a continuous numerical scale allow the observer to capture the variability in quality between different centers (or classrooms) in much greater detail than instruments with closed responses. Despite the clear advantage of a greater variation in the child care quality outcomes, we found, however, no evidence in the literature of a positive relationship between the scoring complexity of an instrument (i.e., continuous numerical scale vs. closed response) and its predictive power for child development outcomes.

It is important to note that training requirements also reflect the type of constructs (or theoretical concepts) targeted by/behind each instrument, as well as the depth with which an instrument seeks to capture each construct, as defined by its manual. We developed a

[17] For example, a trainer and 15 coders/observers participated in this study. ITERS-R training consisted of five days of a train the trainer program at the National Institute for Early Education Research (NIEER) in New Jersey, US, divided into one day of theory and four days of practice to achieve reliability. In the case of the MITRCC, two days of total training time were required to reach reliability among observers. Training for the CC-IT-HOME consisted of four days of a train the trainer program at NIEER, divided into a half day of theory and three and a half days of practice to achieve reliability.

theoretical framework that included all four instruments and identified which process quality constructs (specifically those related to child-caregiver interactions) were measured by each instrument's subscales, as well as by identifying cross-cutting constructs across subscales of these four instruments. We based our theoretical framework on a model developed by Halle, Anderson, Blasberg, Chrisler, and Simkin (2011), who conducted a review of the parent-child and caregiver-child interaction literature. They identified thirteen different types of constructs for child-caregiver interaction covered in this literature, which they grouped into those that measure positive, neutral or negative interactions.[18] Based on this list of constructs, we reviewed the content of each subscale of all four instruments at the item level to determine which child-caregiver construct(s) were covered by each subscale. This mapping exercise is reported in Table 2[19] and shows that the subscales of the four instruments capture a similar set of child-caregiver interaction constructs, with two exceptions: (i) three of the eight positive interaction constructs are not included in the CC-IT-HOME (Support for Peer Interaction, Mutuality, and Joint Attention); and (ii) only CLASS and CC-IT-HOME assess negative interaction.[20] With regard to the relation between these constructs and child outcomes, Halle et al. (2011) found evidence of an association between child outcomes and all of the constructs mentioned in Table 2, with the exception of Support for Peer Interaction and Detachment, which were not examined as predictors of child outcomes in any of the studies. In addition, the authors noted that the strength of the association varied from one construct to another, although a direct comparison across constructs was limited by the fact that the child outcomes and quality measures, as well as the estimation techniques, differed greatly between the various studies. It is important to emphasize, however, that some of these constructs are more prevalent than others in the literature. For instance, of the 35 U.S. studies Halle et al. (2011) reviewed, 18 included Sensitivity/Responsiveness as a construct in their analysis, while only 2 looked at Joint Attention (of which only 10 out of 18 for Sensitivity/Responsiveness looked specifically at these constructs' associations with child outcomes, and 1 out of 2 for Joint Attention).

[Table 2. Child-caregiver interaction constructs]

Furthermore, it is important to note that while there are subscales on different instruments that measure *a priori* similar constructs, the depth with which an instrument seeks to capture the same construct differs between them. For example, while some subscales focus on a tally of certain actions by the caregiver (e.g., the number of times the caregiver performs an action with a child/the children), others require finer observation and attempt to assess whether these were quality actions. This point is illustrated by the Listening and Talking subscale on the ITERS-R and the Caregiver Responsivity subscale on the CC-IT-HOME (whose constructs are notably those of Responsivity as well as Language and Cognitive Stimulation). The former requires the interviewer to evaluate indicators that are more abstract or difficult to observe and tease out, such as "Staff use a wide range of simple exact

[18] The definitions of the constructs are reported in Table A3 of the annex, which corresponds to Table 1 of Halle et al. (2011). There are a number of other dimensions of quality that are not present in the current analysis of the instruments, such as safety, health and nutrition among others, which are included, for instance, in items and indicators in the ITERS-R. However the present analysis focuses on process quality and therefore the constructs in Halle et al. (2011) seem to be the most relevant ones.

[19] We have contacted the publishers of each instrument to get their feedback on our analysis. Table 2 reports the final analysis after having received feedback from all four publishers.

[20] The ITERS-R only partially measures negative interaction constructs (at the level of some groups of indicators).

words in communicating with children," or "Staff are skillful at interpreting children's attempts to communicate and frequently follow through appropriately,"[21] while the latter only requires the interviewer to tally the number of actions, such as "Caregiver spontaneously speaks/vocalizes to child at least twice," or "Caregiver caresses or kisses child at least once."[22] CLASS concepts are even more complex than those of the ITERS-R, which is likely the reason for mandatory training.

The complexity of the instrument is also reflected in the minimum amount of time required for administration. The CC-IT-HOME, for example, can be administered in the field in just one hour, while a minimum of two hours (i.e., four 20-minute observation cycles and 10 minutes of coding for each cycle) is necessary for CLASS administration. A total of three and a half hours is required for the ITERS-R. The brief administration time for the CC-IT-HOME is related not only to the reduced complexity of the constructs, as mentioned previously, but also to the smaller number of items (e.g., the Caregiver Responsivity subscale of the CC-IT-HOME contains just 11 items as compared to the 29 items on the Listening and Talking subscale of the ITERS-R). In addition, a particular feature of the CC-IT-HOME inventory is that the unit of observation is not a group of children but rather the individual interaction between a child and his or her caregiver, which also helps explain its brief administration time. Despite having only 20 binary response (yes/no) items, the MITRCC requires both a relatively long period of observation (three hours) and a higher ability of the observer in order to capture interactions and specific activities, due to the complexity of the underlying constructs for each item (e.g., object permanence and relational correspondence which may also take a long time to occur in a classroom).

Lastly, Table 1 shows that, as expected, although CLASS is capable of capturing the most relevant data with generally higher accuracy, it also has the highest administration costs. For example, its total cost is about 40 times greater than that of the ITERS-R and 20 times greater than that of the CC-IT-HOME.[23, 24]

Procedures

We divide our analysis into three parts. Firstly, we present descriptive statistics of the scores of the four instruments at the subscale level, in order to provide an overview of the process and structural quality of the centers in our sample.

Secondly, since, to our knowledge, there are no other published studies of daycare quality measures for children aged 0-3 in Ecuador, we assess the psychometric properties of the four child care quality instruments, including indicators for internal consistency and reliability, goodness of fit to the publisher's measurement model, construct validity, and concurrent validity. We explore the internal consistency of the instruments using two types of indicators—the correlation between each subscale and the overall score of the same

[21] Note that here the evaluator must be trained on how to observe the concepts of "wide range," "simple exact words," "attempts to communicate," etc.
[22] Still, the correlation between these two subscales is high (0.56, not reported in this text). We intend to explore in future work whether concepts that are easily quantifiable might be strongly correlated to hard to measure quality aspects.
[23] This cost corresponds to the cost of materials per observer-coder. The only exception is CLASS, for which the cost of mandatory official training for observers is included. The time spent by observers to achieve reliability is not taken into account.
[24] Even if the CC-IT-HOME is shorter, the price of the materials is higher than those of the ITERS-R due to both a more expensive manual and a charge for scoresheets which are free for the ITERS-R

instrument, indicating to what extent each of the instrument's subscales are associated with the overall quality measurement that it captures, and Cronbach alphas, representing the degree to which the items in each subscale seem to measure the same concept. Center separation reliability was also estimated for MITRCC using the Rasch model.[25]

We then use CFA to verify that the subscale structure of each instrument conformed to the hypothesized structure set forth by the instrument developers.[26] Measurement models were defined *a priori* for each of the four quality measures and CFAs conducted using the structural equation modeling (SEM) package in Stata/SE version 14.0 (StataCorp LP, College Station, TX, USA). Standardized estimates of the factor loadings and variances were obtained using maximum likelihood estimation and are shown in path diagrams. We assessed the fit for three of the measurement models using the chi-square statistic, the root mean square error of approximation (RMSEA), and the comparative fit index (CFI). The MITRCC model is exactly identified with only three subscales, so the CFA model fit could not be tested. Therefore, we assessed the psychometric properties of the MITRCC using Item Response Theory (IRT), fitting the data to the Rasch model for dichotomous items. To test how well the item responses fit the model, we used the weighted mean square fit statistic, or infit, which is a ratio of the variances of the observed residuals over expected residuals for the model.[27] We identified items where the empirical data deviated significantly from the modeled fit of the data and presented these graphically using item characteristic curves (ICC). We used ACER ConQuest version 2.0 for the IRT models.

In this second section, we also present a series of correlation coefficients and their statistical significance as evidence for both construct and concurrent validity of the instruments. The raw total scores from each instrument are used to generate Pearson correlation coefficients.[28, 29] Evidence of construct validity is demonstrated by correlating each of the total scores with various quality-related factors that we would expect to be related to the instrument scores. These factors are the child-caregiver ratio, location of the center in an urban vs. rural area, proportion of indigenous children in each group, cost of more than or equal to $5 per month to families for their child to attend the center, and the center having been closed since the data collection exercise due to a violation of the INFA quality standards. Evidence of concurrent validity is demonstrated by correlating each of the total scores with each other.

In the third and final section, we evaluate how well each of the subscales of the instruments was able to capture domains of process quality in an Ecuadorian daycare setting. Initially,

[25] High center separation reliability signifies that there is sufficient differentiation in quality scores to distinguish between centers.

[26] We know of one study of the ITERS-R conducted for Chile (Herrera et al., 2005) which looks at the Cronbach's alpha (at the subscale and item level), and the Pearson correlation coefficients with the Caregiver Interaction Scale (Arnett, 1989). CFA was not conducted in this study.

[27] An infit equal to one indicates that the observed residuals vary as much as would be expected, infit values above one denote positive misfit, or more variation than expected, and infit values of less than one denote negative misfit, or less variation than expected. We considered the infit to be acceptable if it fell between 0.75 and 1.33.

[28] A robustness test was performed by analyzing the ranking produced by sorting the centers from worst to best according to their score. The results of the correlations between subscales did not change; however, the correlation between overall scores and CLASS (Table 7) was somewhat lower than the correlation between rankings and CLASS (see Table A6 in the annex).

[29] Correlations are considered very high when in the range of 0.80 to 1, high in the range of 0.6 to 0.8, moderate in the range of 0.4 to 0.6, low in the range of 0.2 to 0.4, and very low when less than 0.2.

we looked at the Pearson correlation of the subscales from the simpler measures (ITERS-R, CC-IT-HOME, and MITRCC) with each of the dimensions of the CLASS.[30] As discussed above, we chose the CLASS as the standard for comparison because, in addition to having the most complex administration, it currently seems to be the measure that bests captures process quality in terms of child-caregiver interactions (Mashburn et al., 2008).

Prior to running the correlations, and based on the analysis presented in Table 2, we hypothesized which of the simple measures' subscales would have the highest number of positive and significant correlations with the CLASS. That set of hypotheses is presented in Table A4 in the appendix. We expect a positive and significant correlation between two subscales when they were shown to have one or more matching constructs in Table 2 (denoted by "+" in Table A4 and Table 8).

Next, using CFA again, we evaluate ten measurement models from our theoretical framework, which links the subscales of all instruments to the child-caregiver interaction constructs for quality interaction that were presented in Table 2, where the instrument subscales where mapped *a priori* to one or more of thirteen child-caregiver interaction constructs. We looked at the factor loadings of the subscales and tested the goodness of fit of the measurement models. Models for three latent constructs: Detachment, Negative Regard and Negative Affect, could not be tested as they were mapped to only two subscales each.

4 Results

Descriptive statistics

Table 3 summarizes the mean scores of these instruments, as well as the scores for the 10th and 90th percentiles. The CLASS utilizes a 7-point scale for scoring each subscale and the total scale, with scores of 1 or 2 indicating low quality; 3 to 5, medium quality; and 6 or 7 indicating high quality. Scoring on the ITERS-R subscales and total scale is based on a 7-point rating scale, with indicators for inadequate quality (score of 1), minimal quality (3), good quality (5), and excellent quality (7). The MITRCC employs the following quality rating to the total score: 6 or below indicates minimal quality; 7, average quality; 8, above-average quality; and 9 or more, high quality.[31]

[Table 3. Child care quality in Ecuador (scores)]

The results in Table 3 reveal that, for the most part, centers provide low-quality child care. This low quality is particularly observed on CLASS, whose scores at the aggregate level fall within the range of low to medium quality (mean of 2.88, SD=0.42), concentrated almost exclusively in the bottom two-thirds of the distribution, and on the ITERS-R, whose total score does not reach the minimal quality range (mean of 2.08, SD=0.53). These results are also observed at the subscale/dimension level of the two measures. On CLASS, five of the eight dimensions have scores indicating low quality, and for dimensions one through five,

[30] We tested other commands such as the polychoric as sensitivity checks and the results stayed about the same or improved (i.e., higher correlations).
[31] The CC-IT-HOME manual, on the other hand, does not define the center's level of quality based on the score obtained.

which compose the Emotional and Behavioral Support domain—key to children's social-emotional development—most of the centers fall within the medium quality range. On the ITERS-R, six of the seven subscales fail to reach a minimal level of quality; even when including better-quality centers—those that fall in the 90th percentile of the distribution for the total score —a minimal quality level is barely attained (mean of 2.85). In addition, on both instruments, these scores show very little dispersion. Nevertheless, there are two positive exceptions in the case of CLASS and the ITERS-R: (i) the Negative Climate dimension of the CLASS reflects a high level of quality; and (ii) the Interaction subscale of the ITERS-R achieves a mean score beyond the range of inadequate quality, into the range of minimal quality (the only subscale to do so). The Interaction subscale is also the only one for which the best centers in the sample (those in the 90th percentile of the distribution) achieve a score of 5, a level of quality considered good (13% of the centers have a score equal to or higher than 5 for this subscale). In contrast, the average scores on the subscales of the CC-IT-HOME and MITRCC, though they also reflect a generally low level of quality, turn out to be slightly higher and more scattered than the scores of the other two instruments.

Psychometric properties

To assess the psychometric properties of the four child care quality instruments, we present evidence for internal consistency and reliability, goodness of fit to the publisher's measurement model, construct validity, and concurrent validity. Table 4 summarizes the two measures of internal consistency mentioned in section 3 for the four instruments.

[Table 4. Internal Consistency]

The results show high Cronbach alphas for all subscales on the four instruments, with the exception of the Personal Care Routines subscale on the ITERS-R, although the manual did note that it yielded a low Cronbach alpha in other studies. The MITRCC has a reasonable alpha at 0.59.[32] The correlation coefficients between each subscale and the instrument's overall score are at least 0.6 for the ITERS-R and CLASS (except the Negative Climate and Regard for Child Perspectives dimensions of CLASS).[33] The CC-IT-HOME exhibits good internal consistency with respect to the Cronbach alphas; however, the correlation coefficients reveal somewhat more variable results, with the Acceptance, Organization, and Variety of Stimulation subscales being particularly low.[34]

The theoretical structure for each of the four quality measures is depicted as a separate path diagram in Figure 2. The latent constructs for each measure are drawn as ovals, the measurement subscales are represented as squares, and factor loadings, or path coefficients, are shown with an arrow, or path, emanating from the latent construct to the subscales. A single latent construct model is used for testing the ITERS-R, CC-IT-HOME and MITRCC, in keeping with the developers' theoretical structure for these measures

[32] Note that the alpha goes up to 0.64 when excluding the three items of the physical development group.
[33] Generally, the literature considers a measure of internal consistency to be reasonable when it falls within the range of 0.6 to 0.7 for both indicators.
[34] Inter-Rater Reliability (IRR) (i.e., the degree of agreement among raters) was also measured during the pilots. CLASS IRR was 90% in all videos and all coders. CC-IT-HOME's IRR was 88% for all coders and ITERS-R 86% for all coders. The lowest IRR is for MITRCC in which the average IRR was 56%.

(Harms, Cryer and Clifford, 2006; Bradley, Caldwell and Corwyn, 2003; MU Center for Family Policy and Research, 2003). The CLASS, on the other hand, is tested with two correlated domains, Engaged Support for Learning and Emotional and Behavioral Support, as described previously (La Paro et al., 2011).

[Figure 2. Confirmatory Factor Analysis for Publishers' Framework, Path diagrams]

Factor loadings for the subscales of the four measures are all statistically significant (all p-values<0.001). For CLASS, the factor loadings for the three subscales on Engaged Support for Learning, as well as for three of the five subscales on Emotional and Behavioral Support, are excellent (0.83 to 0.88). The loading for Regard for Child Perspectives is good at 0.58, but only moderate for Negative Climate at 0.40. The correlation (or standardized covariance) between the two latent constructs is high at 0.85. For ITERS-R, the factor loadings for three of the subscales: Listening and Talking, Interaction, and Program Structure, are very good (0.76-0.85); while the other four are moderate to good (0.49-0.62). The factor loadings for the CC-IT-HOME are moderate to good (0.47-0.71) with the exception of the Acceptance subscale, which loads poorly onto the construct (0.27). As with Negative Climate and CLASS, the Acceptance subscale in the CC-IT-HOME demonstrated very little variation with 74% of the sample achieving a perfect score. For the MITRCC, the factor loadings for the group of items related to physical development was good at 0.52 and excellent for the cognitive group of items at 0.81 (the socio-emotional group of items was arbitrarily chosen to be constrained in our SEM).

The goodness of fit statistics for the models for CLASS, ITERS-R and CC-IT-HOME (Table 5) are all reasonable and similar, although the RMSEA greater than 0.10 for CLASS and ITERS-R suggests that the overall fit to the models is moderate.[35]

[Table 5. Confirmatory Factor Analysis for Publishers' Framework, Goodness of fit statistics]

All but one item out of 20 in the MITRCC fell within the bounds for acceptable fit to the Rasch model. Responses to item 14 in the cognition subscale ("Caregivers talk with children about the beginning and ending of the events of their routines, and help children to anticipate these daily events telling them what is going to happen next") showed unexpected results that were inconsistent with their overall score on the measure, suggesting that some respondents may not have understood the question (infit = 1.36). Item characteristic curves

[35] After testing the fit of the models, we explored other potential model structures based on modification indices for omitted paths (i.e., testing whether the model's goodness-of-fit would improve if a path was added). Results are shown in Table A5 in the appendix. The model fit for the CLASS is improved with the inclusion of a path between Engaged Support for Learning and the Behavior Guidance Subscale, as well as the inclusion of correlated errors between subscales 6 and 8, and 7 and 8 (upper bound of RMSEA = 0.08 and CFI = 0.98). For the ITERS-R, we hypothesized that the subscales are reflecting at least two quality constructs in our data: structural and process. Therefore, we used a combination of experience, expert opinion and factor analysis to explore a two-factor model for the ITERS-R. Specifically, we mapped subscales 1, 4 and 7 to the structural domain, and subscales 2, 3, 5 and 6 to the process domain. The likelihood ratio test indicates that a two-factor model is a better fit to the data than a single-factor model, the factor loadings are higher overall, and the structural equation model fit statistics are also improved (upper bound of RMSEA = 0.08 and CFI = 0.97). We conducted a similar analysis for the CC-IT-HOME, mapping subscales 3 and 4 to the structural domain, and subscales 1, 2 5 and 6 to the process domain. The likelihood ratio test shows that the two-factor model is not a better fit to the data, and the factor loadings for that model are generally lower than for the one-factor model (except for subscales 4 and 5).

for items 14 and 8 are shown in Figure 3, where the empirical data (dashed line) deviate from the modeled fit (solid line) for item 14 (blue), but follow very closely for item 8 (green) that had a very good fit statistic.[36]

[Figure 3. Item characteristic curves, MITRCC (items 14 and 8)]

As evidence of construct validity for the instruments in the Ecuadorian context, we explore the correlations between the instruments' scores and socioeconomic variables, as well as two proxies of center quality (Table 6). Since no household questionnaire was administered to the families of the children in the daycare centers, the socioeconomic indicators that were available in the data were limited to the geographical location of the center (urban or rural), the proportion of indigenous children in each group, and whether families paid a subscription fee equal to or higher than $5 to the center. In addition, in 2015 we asked the INFA to provide us with the most updated list of CIBV centers that were still operating. It turned out that 166 of the centers in the sample (41%) had been closed since the data collection exercise, due to a violation of the INFA quality standards. We use an indicator of whether the center was closed in 2015, and a child-caregiver ratio above the sample median of 9.6[37] as two proxies of center quality.[38] Overall, the correlations are of the expected sign and magnitude. For instance, results show that correlations are positive and significant for centers in urban areas and where families pay a subscription fee equal to or higher than $5. On the other hand, they are negative and significantly correlated for groups with a higher proportion of indigenous children, those which were closed in 2015 and have a child-caregiver ratio above the sample mean. With regard to the magnitude of the coefficients, they are the highest for the subscales of CLASS and ITERS-R, with the highest coefficients being around 0.2 and significant at the 1% level.

[Table 6. Correlations between instruments and sociodemographic indicators]

In terms of concurrent validity, the correlations between the overall CLASS score and the simple measures are low but statistically significant (Table 7).[39] Furthermore, there is little variability in the correlations, with coefficients ranging between 0.32 and 0.36. However, the correlation between the ITERS-R and CC-IT-HOME is very high at 0.80. This is expected given that both measure some structural quality indicators even though their focus is on process quality, and both combine observation and survey methods. Furthermore, their field observer profile is identical. The correlations of the MITRCC with the ITERS-R and CC-IT-HOME are moderate (0.46 and 0.51 respectively).

[36] As an additional check of the performance of the CC-IT-HOME and MITRCC in the Ecuadorian daycare setting, we ran a Rasch model analysis of the individual items. All CC-IT-HOME items fell within the bounds for acceptable fit. The center separation reliability obtained from Rasch model analyses for the CC-IT-HOME and the MITRCC were both very good at 0.84 and 0.86, respectively.

[37] This corresponds to the median using the data collected by the IDB. The median in the INFA administrative data was 9.2. We would have liked to use a lower child-caregiver ratio for this exercise (the literature usually recommends a ratio no higher than 6). However, only 11 centers in our sample have a ratio lower than or equal to 6, 34 have a ratio lower than or equal to 7, and 93 lower than or equal to 8.

[38] We consider a center to be closed when it is no longer in the most updated INFA administrative data. Araujo et al. (2015) show that the centers whose agreements with the INFA had terminated in 2015 were those that in 2012 had, on average, lower levels of quality.

[39] Araujo et al. (2015) also showed that the correlation between CLASS and other measures of quality such as the Knowledge of Infant Development Inventory (KIDI), the Child Care Practices Scale from the National Association for the Education of Young Children (NAEYC), among others, were very low.

[Table 7. Correlations between the CLASS, ITERS-R, CC-IT-HOME and MITRCC (total scores)]

Associations with Process Quality

In order to evaluate how well each of the subscales of the instruments was able to capture process quality, we first looked at the correlations between the simpler subscale scores (i.e., the seven ITERS-R subscales, the six CC-IT-HOME subscales, and the three MITRCC domains) with those on the complex measure (i.e., the eight CLASS dimensions) at the subscale level (see Table 8). [40]

[Table 8. Correlations between CLASS and simple measures, by subscale]

It is particularly interesting to examine the significance and magnitude of correlations in the case of matching constructs (denoted by "+" in Table 8), as well as those unexpected cases in which correlations exist in the absence of matching constructs.

To determine which subscale is most strongly correlated with CLASS, we tallied the number of correlations greater than 0.40. The Listening and Talking subscale of the ITERS-R appears to be the only simple instrument subscale that moderately correlates with the three CLASS dimensions belonging to the Emotional and Behavioral Support domain (Positive Climate, Teacher Sensitivity, and Behavior Guidance), presenting correlations equal to or greater than 0.40. This same subscale also presents a correlation of 0.39 with the CLASS dimension of Language Modeling, which is the most similar construct between the two instrument subscales, thereby showing consistency in measuring the constructs on both instruments. Regardless, we expected a stronger correlation with Facilitation of Learning and Development and Quality of Feedback, a matching construct, and a much weaker correlation with Behavior Guidance, a non-matching construct. The other two instruments do not have correlations in the moderate range with the CLASS dimensions.

Next, we tallied the number of correlations with values between 0.30 and 0.40 for the ITERS-R to determine the second most strongly correlated subscale with CLASS. We found that the ITERS-R subscales of Interaction and Personal Care Routines have an average correlation coefficient of 0.35 again with the three CLASS dimensions belonging to Emotional and Behavioral Support. Except for Behavior Guidance, which measures a different construct, these are precisely the ITERS-R subscales that we would expect to be more highly correlated with CLASS, given the process-focused construct behind them (Harms et al., 2006). It is important to note that neither of the subscales correlated with the dimensions belonging to the Engaged Support for Learning domain, the dimensions most strongly correlated with child development (Berlinski and Schady, 2015). In particular, we expected the Interaction subscale to correlate more strongly with Facilitation of Learning and Development and Quality of Feedback.

[40] Correlations between ITERS-R subscales and the rest of the MITRCC and CC-IT-HOME subscales are discussed in another (unpublished) paper. Readers interested in this analysis can contact the authors at florencial@iadb.org.

As mentioned above, not a single moderate correlation is found between the CC-IT-HOME and CLASS, a surprising result given that, of the three simple measures, the CC-IT-HOME most strongly correlates with CLASS at the aggregate level (0.36). The strongest correlations found are those of Caregiver Responsivity with the Positive Climate and the Teacher Sensitivity dimensions of CLASS (0.31 and 0.32, respectively). Some of the subscales that did not have matching constructs with the CLASS dimensions as per Table 2, do exhibit correlations between 0.20 and 0.30: subscales 3, 4 and 5 of the CC-IT-HOME with Positive Climate; subscales 3, 4 and 6 with Teacher Sensitivity; subscales 3 and 4 with Behavior Guidance; and subscales 5 and 6 with Facilitation of Learning and Development.

For the MITRCC, the highest correlations are found, as expected, between subscales with matching constructs. The highest ones (between 0.30 and 0.40) are those of the socio-emotional group of items with Positive Climate, Teacher Sensitivity and Behavior Guidance, and the cognitive group of items with Behavior Guidance. As expected, the physical development group of items exhibits very low correlations with the CLASS dimensions.

On the other hand, we found that the CLASS dimensions of Negative Climate and Regard for Child Perspectives are very weakly correlated with almost all of the simple measures' subscales, while Quality of Feedback presents very weak to weak correlations with the rest of the subscales. Surprisingly, Negative Climate correlates more strongly with Listening and Talking (0.24) than with the Acceptance subscale of the CC-IT-HOME (0.12). A stronger correlation with the latter was expected, given that it is the only simple instrument subscale that measures negative interactions. In the case of Regard for Child Perspectives, this dimension was expected to correlate with the ITERS-R subscales of Personal Care Routines and Program Structure; it does, in fact, most strongly correlate with these two subscales (0.22 and 0.20, respectively), indicating consistency in measuring constructs.

Finally, we use CFA to evaluate measurement models from our theoretical framework linking the subscales of all instruments to ten of the child-caregiver interaction constructs. Factor loadings and fit statistics for the CFA are presented in Table 9.

[Table 9. Confirmatory Factor Analysis for child-caregiver interaction constructs]

Overall, factor loadings are moderate to excellent (and are all statistically significant at $p<0.001$), with the exception of subscale 7 of the ITERS-R (Parents and Staff) with latent construct Mutuality, and subscale 2 of the CC-IT-HOME (Acceptance of Child), which loads very weakly onto both Sensitivity/Responsiveness and Positive Regard/Warmth. The subscales of the CLASS are those that show the highest loadings, particularly subscales that were mapped to constructs Sensitivity/Responsiveness, Positive Regard/Warmth, and Joint Attention. Subscales 1 (Positive Climate) and 3 (Teacher Sensitivity) display excellent loadings with all their respective constructs (except for subscale 1 with Positive Affect). Loadings for subscales of the other three instruments are moderate to good, with stronger factor loadings onto three theoretical constructs for process quality than did CLASS. For example, the ITERS-R Listening and Talking and Program Structure subscales, as well as the CC-IT-HOME Parental Responsivity and Learning Materials subscales loaded most strongly onto the latent construct for Language & Cognitive Stimulation. The ITERS-R

Interaction subscale, CC-IT-HOME Parental Responsivity subscale, and the MITRCC loaded most strongly onto the latent construct for Positive Affect. And finally, the ITERS-R Interaction subscale, CC-IT-HOME Parental Responsivity and Parental Involvement subscales and MITRCC loaded most strongly onto the latent construct for Behavior Regulatory Style/Guidance. The fit statistics are reasonable for all models, although we observe a better fit for models Positive Affect, Reciprocity, Mutuality and Joint Attention, which show the lowest chi-square values and RMSEA, as well as the highest CFI. The model fit could not be tested for latent construct Intrusiveness as it is exactly identified with only 3 subscales.

6 Conclusion

This study analyzed four instruments that are widely used to measure child care quality for children ages 0 to 36 months, from a sample of 404 child care centers in Ecuador. Our first objective was to test the psychometric properties of the individual instruments. Specifically, we examine their internal consistency, test the underlying subscale structure by means of CFA, verify construct validity by testing associations with quality-related factors (e.g., child-caregiver ratio), and check concurrent validity of the instruments' total scores. Overall, the instruments showed excellent consistency. The CFA also demonstrated that the data were a good fit to the published structure of each instrument (i.e., loadings were high and fit statistics reasonable). Still, we have a few exceptions: Negative Climate for CLASS and Acceptance for the CC-IT-HOME exhibited lower loadings than the other subscales, implying that they may not be a suitable measure of (process) child care quality in the context of the CIBV sample. These findings are very consistent with what we were able to observe in the centers during fieldwork. In fact, the overall level of expressed negativity in the classrooms - the concept these subscales aim to capture - were very low in this specific context. This is also confirmed by the low variability in the data for these two subscales (i.e., 28% of the centers have a perfect score of 7 for Negative Climate, and 73% of the centers have a score of 6 for Acceptance). Associations with quality-related factors exhibited the expected signs, even if there were only a limited number of variables we could use for this exercise. Concurrent validity across instruments revealed low correlations between overall scores on CLASS and the other instruments (0.34 on average) which may be explained by the fact that they are measuring different quality aspects (i.e., CLASS is much more focused on process, and particularly interactions than other instruments). On the other hand, concurrent validity for the other instruments between themselves was very good.

Our second objective in this paper was to use the data on these four instruments to inform and encourage reflection on the development of a simpler and less-expensive monitoring tool. To that end, we firstly explore correlations between the subscales of the simpler instruments with the CLASS dimensions. The following results stand out:

(i) Subscales of the simpler instruments that are more focused on process and interactions showed the highest correlations with the CLASS dimensions, and therefore those worthwhile looking at deeply in future research;

(ii) In particular, the Listening and Talking, Interaction, and Personal Care Routines subscales of the ITERS-R are most strongly correlated with CLASS dimensions (and specifically with the dimensions of Positive Climate, Teacher Sensitivity, and Behavior Guidance, and in the case of the Listening and Talking subscale also with

Facilitation of Learning and Development, Quality of Feedback, and Language Modeling). The first two ITERS-R subscales are precisely those that focus on child-caregiver interactions;

(iii) The other two simple measures yield statistically significant but weaker correlations with CLASS dimensions than the ITERS-R. The subscales with the highest correlations are Caregiver Responsivity from the CC-IT-HOME with the CLASS dimensions of Positive Climate and Teacher Sensitivity, and the items grouped in the social-emotional domain of the MITRCC with the same two CLASS dimensions (as well as Behavior Guidance);

(iv) As expected, very low (and/or insignificant) correlations were found between the CLASS dimension of Negative Climate and the subscales of the simple measures, probably due to the fact that only one of the four instruments measures negative interaction (the Acceptance subscale of the CC-IT-HOME). Contrary to expectations, the correlation between this dimension and the Acceptance subscale of the CC-IT-HOME was very weak. Perhaps in this low-resource context, the ITERS-R subscale is better able to predict this type of interaction.

It is important to note, however, that while these results are overall consistent with our expectations in terms of matching constructs between the instrument's subscales, the magnitude of these correlations is still low (the highest one being around 0.43 - i.e., a moderate correlation – but the other highest correlations ranging between 0.30 and 0.35, which correspond to a low correlation). For instance, the correlations between ITERS-R and CLASS are lower than what were observed in the validation study of CLASS (La Paro et al, 2012).

Lastly, we looked at how the subscales map onto theoretical constructs for process quality and find that the subscales of the CLASS are those that show the highest loadings, particularly subscales that were mapped to constructs Sensitivity/Responsiveness, Positive Regard/Warmth, and Joint Attention. However, several of the subscales of the simpler instruments showed encouraging associations with certain theoretical constructs for process quality that were not as strongly associated with CLASS dimensions. These included The ITERS-R Listening and Talking, Interaction and Program Structure subscales, the CC-IT-HOME Parental Responsivity, Parental Involvement and Learning Materials subscales, as well as the MITRCC. In brief, overall, these constructs seem to be a good base for future research on a simpler measure.

This study has two important methodological advantages. First, to our knowledge, an analysis of such a varied set of center quality measures—administered in the same classrooms, at the same time of day—has never before been performed. The operational design of the fieldwork is unique and fills an important void in the literature, particularly in LAC. Second, the instruments' administration protocols in the field were highly rigorous, generating a high-quality dataset, which allows us to feel confident in the results of this analysis and its recommendations.

In contrast, the main limitation of this analysis lies in the fact that it was conducted in a specific context of low quality, even though we were nonetheless able to exploit a substantial variation in scores within this low quality range. The results found here may differ in a context of higher quality than that of the centers in our sample, which is why we recommend

replicating this study in other contexts to ensure the external validity of the results obtained here. Another limitation of the analysis is that the three simple measures were scored by field observers, while CLASS was scored in a laboratory using filmed classroom observations. Although this method increases the quality of CLASS data, it also reduces the comparability of the scores across the four measures.

In terms of programmatic lessons, these results suggest the possibility of combining some of the simpler instruments subscales or even developing a new instrument that includes the ITERS-R subscales of Listening and Talking, Interaction, and Personal Care Routines, as well as items from CC-IT-HOME to be piloted as a monitoring tool for programs at scale. Such pilot would be administered simultaneously with CLASS and a measure of child development, in order to evaluate whether the newly developed tool does in fact correlate with process quality and is also a predictor of child outcomes. The results in the present analysis also show that if a program's objective were to measure negative constructs, subscales/items of the Acceptance subscale of the CC-IT-HOME (i.e., the only subscale of the simpler instruments that aims to measure negative interactions, yet one with very weak correlation with CLASS) would have to be reformulated in order to correlate more strongly with the CLASS Negative Climate dimension. CLASS remains a relatively expensive instrument, and it has the second longest administration time of the measures analyzed here, rendering it impractical for monitoring on a large scale. An alternative would therefore be to build a new simplified subscale in checklist format on the basis of the Negative Climate dimension of CLASS.

On the other hand, while this analysis is indeed informative for the construction of such a simpler tool, it also highlights two areas of research that require deeper analysis. First, when suggesting that certain ITERS-R subscales can be used to capture process quality, one must take into account that this instrument was not designed with the idea that its subscales would be administered separately and independently from each other. For this reason, it is important to reflect on what the implications would be in terms of administration time, cost, and training for a program that wishes to administer just one (or a few) of this instrument's subscales. Second, the fact that none of the simple instrument subscales yielded a strong (or very strong) correlation with CLASS dimensions suggests the need for an analysis at an even more disaggregated level than the one performed by this study. One possible methodology would involve the comparison of correlations between the *items* of the simple instrument subscales and CLASS dimensions. Another possibility to consider is to further explore these correlations by regrouping items from various subscales (either within the same instrument or across the four instruments) in order to build a simplified tool that is highly correlated with process quality, and that serves as a predictor of child development.

Finally, it is important to remember that, when developing a simple tool that enables the continuous monitoring of child care process quality, one must take into account not only the results presented here in terms of correlations between instrument scores, but also the tradeoffs associated with the uses of each of these subscales. For example, this analysis suggests that certain ITERS-R subscales could be used to capture process quality, given

that the latter is a less complex and costly instrument than CLASS; however, one must take into account that the full instrument still requires three and a half hours of observation, an adequate observer profile,[41] and intensive training. These requirements would remain a major challenge for some programs in the LAC region, and as such recommendations for the creation of a simpler tool should take into account the available resources of countries' existing monitoring systems (such as the financial resources, information systems and training capacity of the monitoring staff).

[41] For example, a study of the MITRCC showed that there are considerable differences between professional and non-professional coders. Future studies should examine the importance of the trainer profile, the time of day that coding is performed, the activity coded, etc. (Hill et al., 2012).

References

Araujo, M. C., López-Boo, F., Novella, R., Schodt, S., & Tomé, R. (2015). "La calidad de los Centros Infantiles del Buen Vivir en Ecuador," Policy Brief Nº PB-248, Washington DC, Inter-American Development Bank.

Arnett, J. (1989). Caregivers in day care centres: does training matter?, Journal of Applied Developmental Psychology, 10, 541–552.

Bradley, R., Caldwell, B., & Corwyn, R. (2003). The Child Care HOME Inventories: Assessing
the quality of family child care homes. *Early Childhood Research Quarterly*. 18: 294-309.

Berlinski, S., & Schady, N. (2015). The Early Years: Child Well-being and the Role of Public Policy. Development in the Americas Series. New York: Palgrave MacMillan, and Washington, DC: Inter-American Development Bank.

Bowlby, J. (1969). Attachment. Attachment and loss: Vol. 1. Loss. New York: Basic Books.

Burchinal, M., Campbell, F., Bryant, D., Wasik, B., & Ramey, C. (1997). "Early intervention and mediating processes in cognitive performance of children of low- income African American families." *Child Development*. 68: 935-954.

Campbell, F., & Ramey, C. (1995). "Cognitive and school outcomes for high-risk African-American students at middle adolescence: Positive effects of early intervention." *American Educational Research Journal*. 32: 743-772.

Greenberg, M., Domitrovich, C., & Bumbarger, B. (2001). "The prevention of mental disorders in school-aged children: Current state of the field." *Prevention & Treatment*, 4, Article 1.

Halle, T., Anderson, R., Blasberg, A., Chrisler, A., & Simkin, S. (2011). Quality of Caregiver-Child Interactions for Infants and Toddlers (QCCIIT): A Review of the Literature, OPRE 2011- 25. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Hamre, B., & Pianta, R. (2007). Learning opportunities in preschool and early elementary classrooms. In R. C. Pianta, M. J. Cox & K. L. Snow (Eds.), *School readiness & the transition to kindergarten in the era of accountability.* 49-83. Baltimore, MD: Paul H. Brookes Publishing Co.

Harms, T., Cryer, T. & Clifford, R. (2006). Infant/Toddler Environment Rating Scale Revised Edition — ITERS-R, Teachers College Press.

Herrera, M. O., Mathiesen, M. E., Merino, J. M., & Recart, I. (2005). Learning contexts for young children in Chile: Process quality assessment in preschool centres. International Journal of Early Years Education, 13, 13–27.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough teacher observation systems and a case for the generalizability study. Educational Researcher, 41(2), 56-64.

Howes, C., Phillips, D., & Whitebook, M. (1992). “Thresholds of quality: Implications for the social development of children in center-based child care.” Child Development. 63: 449-460.

Kane, T. J., & Staiger, D. O. (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project*. Bill & Melinda Gates Foundation.*

Li, W., Farkas, G., Duncan, G., Burchinal, M., & Vandell, D. L. (2013). “Timing of high quality child care and cognitive, language, and preacademic development.” *Developmental Psychology* 49(8) August: 1440-51.

La Paro, K., Pianta, R., & Stuhlman, M. (2004). “The classroom assessment scoring system: Findings from the prekindergarten year.” *Elementary School Journal*.104: 409–426.

La Paro, K., Hamre, B. & Pianta, R.,. (2012). “*Classroom assessment scoring system Manual, Toddle*r” Paul Brookes Publishing Co.

López-Boo, F., Araujo, M. C., & Tomé, R. (2016). *¿Cómo se mide la calidad de los servicios de cuidado?* Guía de herramientas para la medición de calidad de centros de cuidado de bebés y niños pequeños, Washington DC, Inter-American Development Bank.

Mashburn, A., Pianta, R. Hamre, B., Downer, J., Barbarin, O., Bryant D., Burchinal, M., Early, D., & Howes, C. (2008). “Measures of Classroom quality in prekindergarten and children’s development of academic, language, and social skills.” *Child Development*, May/June 2008, vol.79, number 3, pages 732-749.

MU Center for Family Policy and Research. (2003). Missouri Infant/Toddler Responsive Caregiving Checklist. Retrieved from https://www.openinitiative.org/content/pdfs/MoNotes/IT_Checklist_Notes.pdf

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). Classroom assessment scoring system, manual, K–3. Baltimore, MD: Brookes Publishing.

NICHD Early Child Care Research Network. (2000a). “Characteristics and quality of child care for toddlers and preschoolers.” *Applied Developmental Science*. 4: 116-135.

NICHD Early Child Care Research Network. (2000b). “The relation of child care to cognitive and language development.” *Child Development*. 71: 960-980.

Peisner-Feinberg, E., Burchinal, R., Clifford, R., Culkin, M., Howes, C., Kagan, S. & Yazejian, N. (2001). “The Relation of Preschool Child-Care Quality to Children’s Cognitive and Social Developmental Trajectories through Second Grade.” *Child Development.* 72(5): 1534-1553.

Shonkoff, J., & Phillips, D. (Eds.). (2000). From Neurons to Neighbourhoods: *The Science of Early Childhood Development.* National Academy Press, Washington, DC.

Verdisco, A., & Pérez Alfaro, M. (2010). “Measuring Education Quality in Brazil.” Briefly Noted Series No. 6. Education Division, Inter-American Development Bank, Washington, DC. Retrieved from https://publications.iadb.org/bitstream/handle/11319/3100/Measuring%20Education%20Quality%20in%20Brazil.pdf?sequence=1 Vandell, D. L., Belsky, J., Burchinal, M.,

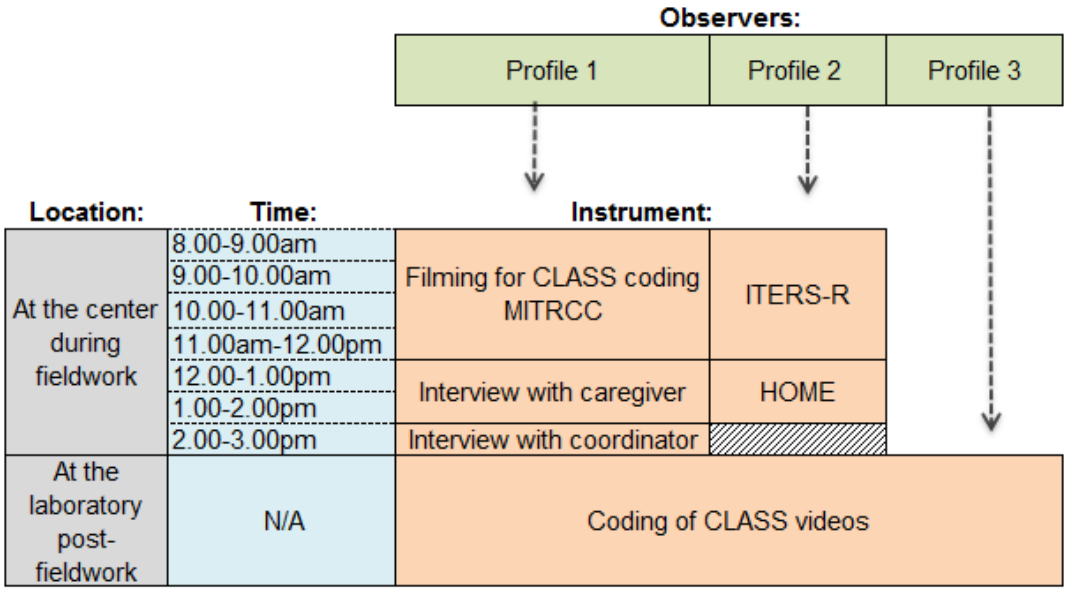
Vandergrift, N., & Steinberg, L. (2010). “Do Effects of Early Child Care Extend to Age 15 Years? Results From the NICHD Study of Early Child Care and Youth Development.” *Child Development,* 81(3) May-June: 737-756.

Vogel, C. A., Caronongan, P., Thomas, J., Bandel, E., Xue, Y., Henke, J., Aikens, N., Boller, K., & Murphy, L. (2015). Administration for Children and Families. *Toddlers in Early Head Start: A Portrait of 2-Year-Olds, Their Families, and the Programs Serving Them.* OPRE Report #2015-10, Washington, DC. Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Zaslow, M., Halle, T., Martin, L., Cabrera, N., Calkins, J., Pitzer L., & Geyelin, N. (2006). “Child Outcome Measures in the Study of Child Care Quality.” *Eval Rev*; 30; 57

Tables and Figures

Figure 1. Protocol for instrument administration/coding in each center



Notes:
Profile 1 = complete secondary education
Profile 2 = post-secondary education in the field of child psychology or early childhood education
Profile 3 = post-secondary education in the field of child psychology or early childhood education

Table 1. Main characteristics of the child care quality measures

	CLASS	ITERS-R	CC-IT-HOME	MITRCC
Age range (months)	15-36	0-30	0-36	0-36
Type of variables	Process	Structural/ process	Structural/ process	Structural/ process
Number of dimensions/subscales	8	7 (1 structural)	6 (1 structural)	N/A
Number of items	8	39	43	20
Assessment method	Observation (direct or video)	Direct observation and reporting	Direct observation and reporting	Direct observation
Official training	Yes	No	No	No
Scoring method	1-7	1-7 *	0-1 (yes/no)	0-1 (yes/no)
Minimum administration time (hours)	2 **	3,5	1	3,5
Total cost (in US$) ***	902,90	22,90	40,30	0,00
Adapted for any country in LAC	Yes	Yes	Yes	No
Spanish translation	Yes	Yes	Yes	Yes
Validity ****	Yes ^^^	Yes ^	Yes ^^	Yes ^^^^

Notes: *ITERS-R scoring is administered gradually, meaning that once a classroom fails to comply with a subset of items within a given subscale, it receives the score corresponding to the highest stop point.**Unlike other instruments, administration begins with classroom video footage of a four-hour day, from which four 20-minute observation cycles are extracted. In addition, CLASS is the only instrument that is not scored in the field and that requires a post-fieldwork phase for the coding of videos. ***This cost corresponds to the cost of materials per observer/coder. The only exception is CLASS, for which the cost of the mandatory, official train-the-trainer program is included. Prices were valid as of December 28, 2015. ****This characteristic indicates the existence of studies on the instrument's validity and reliability (López Boo et al., 2016). To determine an instrument's validity, it is necessary to analyze whether the measure is appropriate for the population in terms of cultural relevance and language or if, instead, translation or adaptation of the measure to the language or local context is required. ^Concurrent and predictive validity; ^^Criterion, construct, concurrent, discriminant, convergent and content validity; ^^^Construct validity (additional validity information forthcoming) ; ^^^^Concurrent and predictive validity.

Table 2. Child-caregiver interaction constructs

		Positive								Neutral	Negative			
		Sensitivity/Responsiveness	Language & Cognitive Stimulation	Support for Peer Interaction	Positive Regard/Warmth	Positive Affect	Reciprocity	Mutuality	Joint Attention	Behavior Regulatory Style/Guidance	Detachment	Intrusiveness	Negative Regard	Negative Affect
CLASS	1. Positive Climate	×		×	×	×	×		×		×			
	2. Negative Climate											×	×	×
	3. Teacher Sensitivity	×			×				×		×			
	4. Regard for Child Perspectives		×		×		×	×				×		
	5. Behavior Guidance									×				
	6. Facilitation of Learning and Development		×					×	×	×		×		
	7. Quality of Feedback		×				×	×						
	8. Language Modeling		×	×										
ITERS-R	1. Space and Furnishings													
	2. Personal Care Routines	×			×	×								
	3. Listening and Talking	×	×											
	4. Activities		×	×			×		×					
	5. Interaction	×		×	×	×				×				
	6. Program Structure		×											
	7. Parents and Staff							×		×				
CC-IT-HOME	1. Caregiver Responsivity	×	×		×	×				×				
	2. Acceptance of Child	×			×	×							×	×
	3. Organization of the Environment		×											
	4. Learning Materials		×											
	5. Caregiver Involvement	×					×			×				
	6. Variety in Experience		×											
	MITRCC (total)	×	×	×	×	×	×	×	×	×				

Source: Own analysis based on Halle, Anderson, Blasberg, Chrisler, and Simkin (2011). See Table A3 in the appendix for the constructs' definitions.

Table 3. Child care quality in Ecuador (scores)

	Mean	SD	Possible range of scores	P10	P90
CLASS					
1. Positive climate	3.34	0.62	[1-7]	2.63	4.13
2. Negative climate	6.61	0.44	[1-7]	6.00	7.00
3. Teacher sensitivity	3.36	0.63	[1-7]	2.63	4.13
4. Regard for child perspectives	1.98	0.28	[1-7]	1.67	2.38
5. Behavior guidance	2.85	0.50	[1-7]	2.25	3.50
6. Facilitation of learning and development	2.08	0.53	[1-7]	1.50	2.75
7. Quality of feedback	1.30	0.33	[1-7]	1.00	1.75
8. Language modeling	1.56	0.51	[1-7]	1.00	2.25
Total score	2.88	0.42	[1-7]	2.44	3.35
ITERS-R					
1. Space and furnishings	2.10	0.62	[1-7]	1.40	3.00
2. Personal care routines	1.69	0.54	[1-7]	1.00	2.50
3. Listening and talking	2.48	1.18	[1-7]	1.00	4.50
4. Activities	1.54	0.47	[1-7]	1.00	2.22
5. Interaction	3.30	1.26	[1-7]	1.50	5.00
6. Program structure	2.57	1.26	[1-7]	1.00	4.33
7. Parents and staff	2.00	0.64	[1-7]	1.17	2.83
Total score	2.08	0.53	[1-7]	1.41	2.85
CC-IT-HOME					
1. Caregiver responsivity	6.68	2.52	[0-11]	3.00	10.00
2. Acceptance	5.69	0.59	[0-6]	5.00	6.00
3. Organization	2.85	0.99	[0-6]	2.00	4.00
4. Learning materials	4.47	2.09	[0-9]	2.00	7.00
5. Caregiver involvement	3.63	1.76	[0-6]	1.00	6.00
6. Variety of stimulation	1.37	0.79	[0-4]	1.00	3.00
Total score	24.69	6.06	[0-42]	17.00	33.00
MITRCC					
1. Social-emotional	2.30	1.57	[0-5]	0.00	4.00
2. Physical development	1.61	0.93	[0-3]	0.00	3.00
3. Cognitive	4.92	2.82	[0-11]	1.00	9.00
Total score	8.78	4.62	[0-20]	2.00	15.00

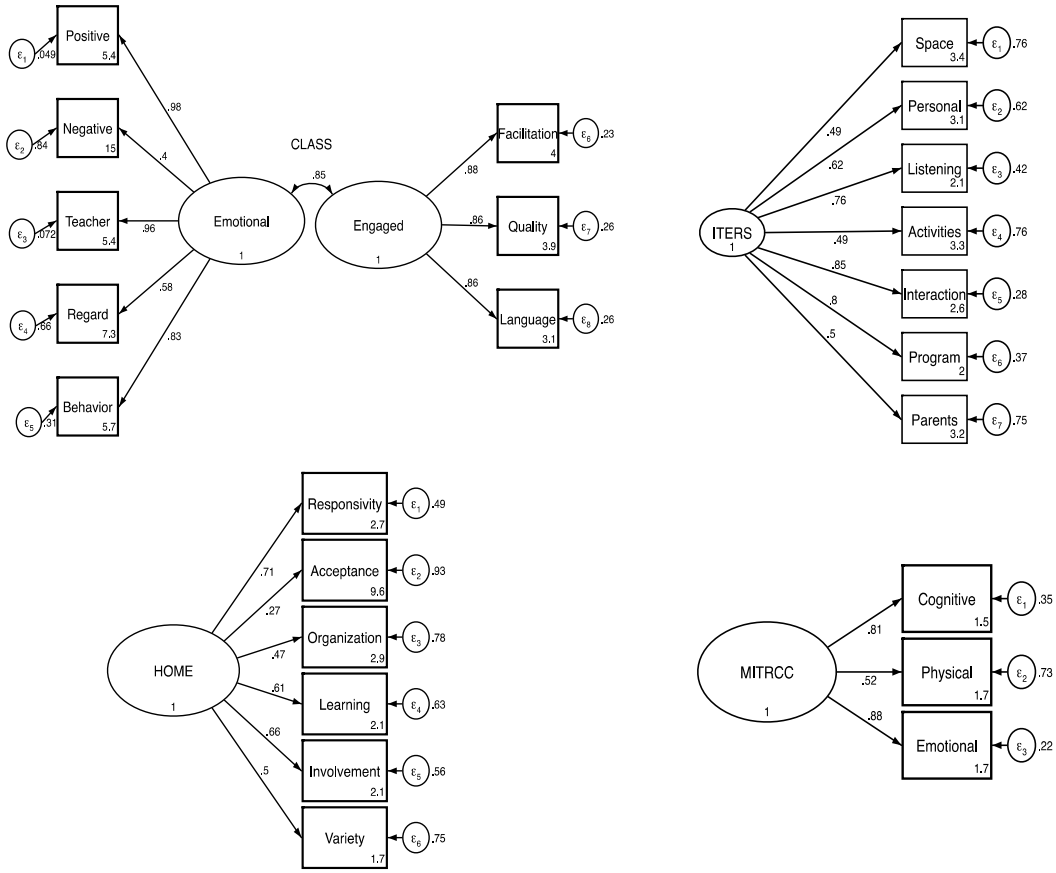
Notes: The data in table 3 correspond to the sample of 404 child care centers. P10 refers to the 10th percentile, and P90 refers to the 90th percentile.

Table 4. Internal Consistency

	Correlation	Alpha
CLASS		
Emotional and behavioral support	0.97***	0.86
1. Positive climate	0.93***	
2. Negative climate	0.50***	
3. Teacher sensitivity	0.92***	
4. Regard for child perspectives	0.59***	
5. Behavior guidance	0.89***	
Engaged support for learning	0.91***	0.9
6. Facilitation of learning and development	0.86***	
7. Quality of feedback	0.80***	
8. Language modeling	0.81***	
Full scale		0.91
ITERS-R		
1. Space and furnishings	0.64***	0.61
2. Personal care routines	0.69***	0.49
3. Listening and talking	0.76***	0.61
4. Activities	0.65***	0.63
5. Interaction	0.82***	0.77
6. Program structure	0.78***	0.76
7. Parents and staff	0.64***	0.57
Full scale		0.85
CC-IT-HOME		
1. Caregiver responsivity	0.83***	0.74
2. Acceptance	0.32***	0.51
3. Organization	0.53***	0.26
4. Learning materials	0.74***	0.66
5. Caregiver involvement	0.74***	0.66
6. Variety of stimulation	0.53***	0.35
Full scale		0.83
MITRCC		
Full scale		0.59

Notes: This table shows the Pearson correlation coefficients of subscale scores with the total score of the instrument and the Cronbach alphas. For the ITERS-R, results do not factor in items 23, 32 and 36 because they applied to less than 15%, 5% and 11% of the centers, respectively. Correlations significant at *** p<0.01.

Figure 2. Confirmatory Factor Analysis for Publishers' Framework, Path diagrams



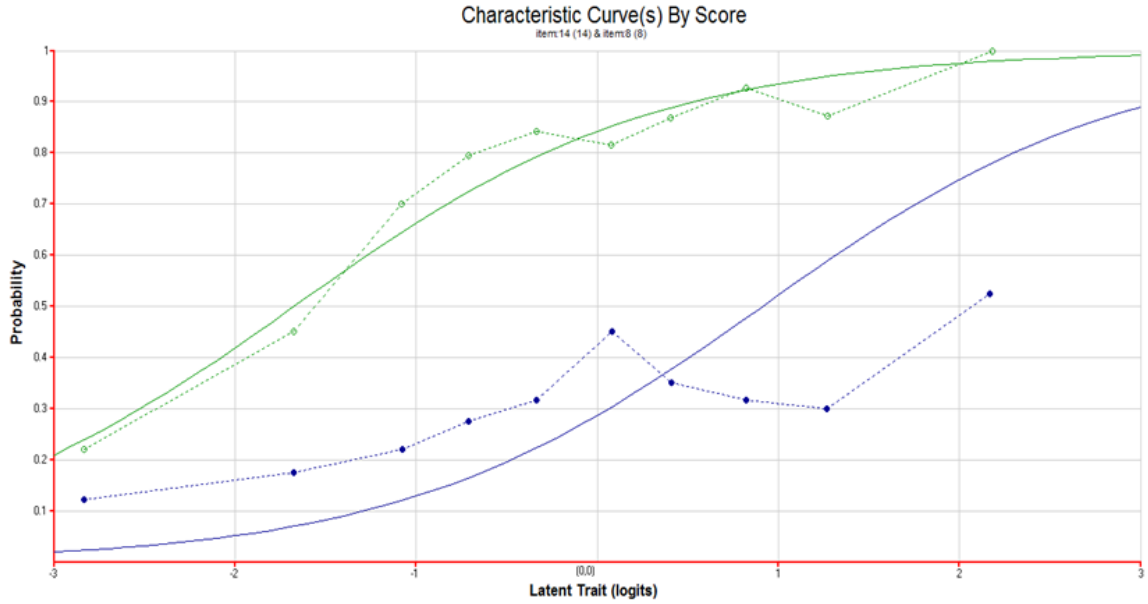
Notes: Confirmatory factor analysis of two-factor model for CLASS, and single-factor models for ITERS-R, CC-IT-HOME, and MITRCC, standardized estimates.

Table 5. Confirmatory Factor Analysis for Publishers' Framework, Goodness of fit statistics

Statistic	CLASS	ITERS-R	CC-IT-HOME
Chi-Square	165,22	118,00	33,32
Df	19,00	14,00	9,00
RMSEA	0,14	0,14	0,08
CFI	0,95	0,90	0,94

Notes: Df refers to the degrees of freedom, RMSEA to the root mean square of approximation, and CFI to the comparative fit index.

Figure 3. Item characteristic curves, MITRCC (items 14 and 8)



Notes: Item characteristic curves for items 14 (blue) and 8 (green) of the MITRCC. The empirical data is shown by the dashed line and the model fit by the solid line.

Table 6. Correlations between instruments and sociodemographic indicators

	Proxies for family socioeconomic status			Proxies for center quality	
	Urban center	Proportion of indigenous children	Families pay monthly fee ≥$5	Closed center	Child-caregiver ratio above median (9.6)
CLASS					
1. Positive climate	0.14***	-0.19***	0.23***	-0.14*	-0.09*
2. Negative climate	-0.03	0.02	-0.01	0.00	-0.10**
3. Teacher sensitivity	0.13***	-0.20***	0.22***	-0.14*	-0.09*
4. Regard for child perspectives	0.04	-0.04	0.04	-0.03	-0.06
5. Behavior guidance	0.13***	-0.19***	0.19***	-0.11*	-0.07
6. Facilitation of learning and development	0.11**	-0.15***	0.09*	-0.11*	-0.12**
7. Quality of feedback	0.09*	-0.16***	0.10**	-0.14*	-0.04
8. Language modeling	0.14***	-0.23***	0.10**	-0.16*	-0.04
Total score	0.09*	-0.16***	0.11**	-0.11*	-0.11**
ITERS-R					
1. Space and furnishings	0.17***	-0.15***	0.13***	-0.13*	0.01
2. Personal care routines	0.20***	-0.17***	0.12**	-0.15*	-0.03
3. Listening and talking	0.08*	-0.08	0.11**	-0.05	-0.02
4. Activities	0.09*	-0.11**	0.16***	-0.04	-0.08*
5. Interaction	0.17***	-0.08	0.09*	-0.11*	-0.04
6. Program structure	0.15***	-0.09	0.08	-0.12*	-0.04
7. Parents and staff	0.15***	-0.08*	0.16***	-0.11*	0.00
Total score	0.20***	-0.13***	0.17***	-0.14*	-0.047
CC-IT-HOME					
1. Caregiver responsivity	0.18***	-0.18***	0.06	-0.11*	0.00
2. Acceptance	0.02	0.02	0.02	-0.07	0.04
3. Organization	0.12**	-0.07	0.13***	-0.05	0.04
4. Learning materials	0.22***	-0.04	0.11**	-0.14*	-0.04
5. Caregiver involvement	0.16***	-0.09*	0.09*	-0.07	-0.14***
6. Variety of stimulation	0.12**	-0.14***	0.03	-0.11*	-0.08*
Total score	0.24***	-0.14***	0.12**	-0.15*	-0.05
MITRCC					
1. Social-emotional	0.15***	-0.08*	0.01	-0.12*	-0.04
2. Physical development	0.12**	-0.03	0.09*	-0.11*	0.02
3. Cognitive	0.15***	-0.12**	0.00	-0.12*	-0.02
Total score	0.17***	-0.11**	0.02	-0.13*	-0.02

Notes: Pearson correlation coefficients. "Families pay monthly fee ≥$5" is a dichotomous variable equal to 1 if families pay a fee greater than or equal to $5 for the child to attend the daycare center (20% of centers). "Closed center" refers to a center that was not found in the 2015 INFA administrative data, and as such we considered closed. Coefficients significant at *p<0.10. **p<0.05 and ***p<0.01.

Table 7. Correlations between the CLASS, ITERS-R, CC-IT-HOME and MITRCC (total scores)

	CLASS	ITERS-R	CC-IT-HOME	MITRCC
CLASS	1			
ITERS-R	0.34***	1		
CC-IT-HOME	0.36***	0.80***	1	
MITRCC	0.32***	0.46***	0.51***	1

Notes: Pearson correlation coefficients. Coefficients significant at * p<0.10,** p<0.05, and *** p<0.01.

Table 8. Correlations between CLASS and simple measures, by subscale

	CLASS							
	1. Positive climate	2. Negative climate	3. Teacher sensitivity	4. Regard for child perspectives	5. Behavior guidance	6. Facilitation of learning and development	7. Quality of feedback	8. Language modeling
ITERS-R								
1. Space and furnishings	0.19***	-0.02	0.20***	0.10*	0.12**	0.10**	0.06	0.13**
2. Personal care routines	0.35***+	0.09*	0.35***+	0.22***+	0.33***	0.26***	0.24***	0.27***
3. Listening and talking	0.41***+	0.23***	0.40***+	0.23***	0.43***	0.35***+	0.30***+	0.39***+
4. Activities	0.25***+	0.11**	0.22***+	0.14***+	0.24***	0.18***+	0.15***+	0.20***+
5. Interaction	0.35***+	0.16***	0.34***+	0.16***+	0.34***+	0.25***+	0.21***	0.27***+
6. Program structure	0.24***	0.11**	0.26***	0.20***+	0.24***	0.18***+	0.13***+	0.16***+
7. Parents and staff	0.11** +	-0.10*	0.10*	0.02	0.08*+	0.03+	0.06+	0.09*
Total score								
CC-IT-HOME								
1. Caregiver responsivity	0.31***+	0.14***	0.32***+	0.11** +	0.28***+	0.23***+	0.22***+	0.27***+
2. Acceptance	0.03+	0.12**+	0.03+	-0.05	0.03	-0.01	0.038	0.04
3. Organization	0.26***	0.12**	0.23***	0.16***+	0.25***	0.19***+	0.18***+	0.19***+
4. Learning materials	0.24***	0.09*	0.24***	0.19***+	0.23***	0.20***+	0.13***+	0.18***+
5. Caregiver involvement	0.24***+	0.08	0.26***+	0.14***	0.25***+	0.23***	0.21***+	0.20***
6. Variety of stimulation	0.25***	0.06	0.23***	0.17***+	0.24***	0.27***	0.21***+	0.29***+
Total score								
MITRCC								
1. Social-emotional	0.31***+	0.09*	0.30***+	0.15***	0.31***+	0.29***	0.24***	0.28***+
2. Physical development	0.14***	0.15***	0.14***	0.05	0.12***	0.08	0.05	0.03
3. Cognitive	0.27***	0.08*	0.26***+	0.18***	0.30***	0.28***+	0.20***	0.22***+
Total score								

Notes: Pearson correlation coefficients. "+" refers to an anticipated positive and significant correlation between two subscales based on the analyses in Tables 2 and A4. Coefficients significant at * p<0.10, ** p<0.05, and *** p<0.01.

Table 9. Confirmatory Factor Analysis for child-caregiver interaction constructs

		Latent Construct								
		Sensitivity/Responsiveness	Language & Cognitive Stimulation	Support for Peer Interaction	Positive Regard/Warmth	Positive Affect	Reciprocity	Mutuality	Joint Attention	Behavior Regulatory Style/Guidance	Intrusiveness
Factor loadings											
CLASS	1. Positive Climate	0.97		0.89	0.97	0.44	0.94		0.98		
	2. Negative Climate										0.41
	3. Teacher Sensitivity	0.97			0.97				0.95		
	4. Regard for Child Perspectives		0.39		0.58		0.60	0.49			0.60
	5. Behavior Guidance									0.46	
	6. Facilitation of Learning and Development		0.54					0.94	0.74	0.40	0.72
	7. Quality of Feedback		0.49				0.71	0.81			
	8. Language Modeling		0.54	0.79							
ITERS-R	1. Space and Furnishings										
	2. Personal Care Routines	0.38			0.37	0.58					
	3. Listening and Talking	0.45	0.76								
	4. Activities		0.51				0.27		0.25		
	5. Interaction	0.39			0.37	0.85				0.81	
	6. Program Structure		0.68								
	7. Parents and Staff							0.05		0.40	
CC-IT-HOME	1. Parental Responsivity	0.35	0.64		0.34	0.75				0.76	
	2. Acceptance of Child	0.05			0.04	0.25					
	3. Organization of the Environment		0.53								
	4. Learning Materials		0.61	0.31							
	5. Parental Involvement	0.28		0.42			0.27			0.63	
	6. Variety in Experience		0.50								
	MITRCC (total)	0.32	0.54	0.36	0.31	0.60	0.33	0.30	0.30	0.61	
Goodness of fit statistics											
	Chi-Square	780.07	865.00	119.62	409.50	26.62	55.56	22.31	21.02	344.91	
	Df	27.00	54.00	5.00	20.00	9.00	9.00	5.00	5.00	14.00	
	RMSEA	0.26	0.19	0.24	0.22	0.07	0.11	0.09	0.09	0.24	
	CFI	0.61	0.59	0.79	0.76	0.97	0.91	0.97	0.99	0.67	
	CD	0.97	0.86	0.86	0.97	0.84	0.90	0.90	0.98	0.84	

Notes: Own analysis based on Halle, Anderson, Blasberg, Chrisler, and Simkin (2011). Df refers to the degrees of freedom, RMSEA to the root mean square of approximation, CFI to the comparative fit index, and CD to the coefficient of determination. The loadings for constructs Detachment, Negative Regard and Negative Affect could not be tested because they are mapped to only two subscales. The Intrusiveness model is exactly identified with three subscales, so the model fit could not be estimated.

Appendix

Table A1. Population and sample characteristics

	Population			Sample		
	Coef.≥ 9.2	Coef.< 9.2	Diff. (t-test)	Coef.≥9.2	Coef.< 9.2	Diff. (t-test)
Child-caregiver ratio	12.10	7.53	***	11.99	7.72	***
Total number of children	36.89	28.37	***	35.92	27.50	***
Children ages 0-1	1.38	1.28	*	1.30	1.25	
Children ages 1-2	5.74	4.68	***	5.77	4.76	***
Children ages 2-3	9.40	7.39	***	9.06	7.34	***
Children age 3 and older	20.37	15.02	***	19.80	14.15	***
Total number of staff	4.85	4.94		4.72	4.81	
Caregivers	3.28	3.79	***	3.19	3.61	***
Food service staff	1.58	1.15	***	1.53	1.20	***
Geographic location						
Urban (%)	53.34	45.44	***	49.00	38.00	*
Canton with provincial capital (%)	41.82	38.40	*	38.00	42.00	
Coast (%)	36.71	26.91	**	35.00	22.50	***
Mountains (%)	50.42	59.35	***	53.50	63.50	**
Amazon region (%)	12.87	13.74		11.50	14.00	
Type of operating entity						
Municipality (%)	25.69	32.55	**	25.00	37.00	**
Parish council (%)	22.37	24.04		26.00	22.50	
Intl. NGO/committee/religious entity (%)	11.69	8.22	**	9.00	10.00	
Provincial/central government (%)	0.45	0.68	*	0.00	1.50	
Others (%)	37.32	31.81		38.00	28.00	*
Observations	1,779	1,776		201	203	

Notes: The median child-caregiver ratio in the administrative data is 9.2. The sample was stratified in two groups of centers: those with 9.2 children per adult or more, and those with less than 9.2 children per adult. Standard errors of the mean adjusted for clustering at the canton level. Differences between the two groups significant at *p<0.10, **p<0.05 and ***p<0.01.

Table A2. Instrument scores, by level of child care quality

	Low quality		High quality		P-value of difference	
	Mean	SD	Mean	SD		
CLASS						
1. Positive climate	3.28	0.66	3.40	0.58	0.10	*
2. Negative climate	6.57	0.50	6.66	0.37	0.02	**
3. Teacher sensitivity	3.31	0.64	3.42	0.61	0.11	
4. Regard for child perspectives	1.97	0.30	2.00	0.25	0.24	
5. Behavior guidance	2.82	0.52	2.89	0.48	0.26	
6. Facilitation of learning and development	2.02	0.53	2.14	0.52	0.03	**
7. Quality of feedback	1.29	0.32	1.31	0.34	0.49	
8. Language modeling	1.54	0.51	1.58	0.51	0.43	
Total score	2.83	0.54	2.93	0.38	0.03	**
ITERS-R						
1. Space and furnishings	2.11	0.60	2.10	0.64	0.79	
2. Personal care routines	1.67	0.51	1.71	0.57	0.45	
3. Listening and talking	2.46	1.20	2.51	1.16	0.66	
4. Activities	1.50	0.43	1.58	0.50	0.10	*
5. Interaction	3.25	1.19	3.35	1.32	0.44	
6. Program structure	2.51	1.18	2.62	1.33	0.41	
7. Parents and staff	2.00	0.67	2.01	0.61	0.94	
Total score	2.06	0.49	2.11	0.57	0.35	
CC-IT-HOME						
1. Caregiver responsivity	6.69	2.48	6.67	2.56	0.93	
2. Acceptance	5.72	0.51	5.66	0.67	0.43	
3. Organization	2.90	0.95	2.81	1.02	0.32	
4. Learning materials	4.39	1.99	4.55	2.18	0.48	
5. Caregiver involvement	3.37	1.78	3.87	1.71	0.01	***
6. Variety of stimulation	1.30	0.75	1.44	0.83	0.12	
Total score	24.37	5.79	25.00	6.33	0.33	
MITRCC						
1. Social-emotional	2.25	1.61	2.36	1.54	0.53	
2. Physical development	1.63	0.93	1.58	0.93	0.61	
3. Cognitive	4.84	2.85	4.95	2.81	0.73	
Total score	8.72	4.72	8.89	4.51	0.74	

Notes: the sample is divided into two groups of centers: those with child-caregiver ratios above the sample median of 9.6 (low-quality centers), and those with child-caregiver ratios below the sample median (high-quality centers). The last column shows the p-value for the null hypothesis that the difference between the scores of the two groups is equal to zero. The study's pre-survey stratification was performed using the INFA data reported in Araujo et al. (2015) where the median child-caregiver ratio was 9.2. Coefficients significant at * p<0.01, ** p<0.05, and *** p<0.10.

Table A3. Definitions and examples of child-caregiver interaction constructs

Interaction Construct/Definition	Examples
Positive Interaction Constructs	
Sensitivity/Responsiveness: Responding to the needs of individual children and acknowledging children's feelings and thoughts	"Provider is attentive and responsive to the children" (APFCCH). "Provider regularly responds contingently to children's questions and queries in ways that support children's activity" (CHELLO). "Teaching staff is flexible and responsive in interaction with children" (CDPES). "Teacher responds to infant's physical gestures" (APECP).
Language & Cognitive Stimulation: Providing opportunities for children to develop language through conversation and providing opportunities for children to develop cognitive skills through activities	"Provider regularly encourages children's verbal interactions by asking questions, encouraging elaborations, and supporting continual exchanges" (CHELLO). "Adds to children's attempts to dialogue; adds words and explanations to talk" (CCIS). "Staff talk with children about ideas related to their play (for example, bring in concepts such as near-far, fast-slow for younger children; ask children to tell about building project or dramatic play)" (ECERS-R).
Support for Peer Interaction: Providing support for and prompting children to interact with one another	"Encourages children to exhibit pro-social behavior, e.g. sharing, helping" (CIS). "Teacher teaches children about sharing, taking turns, and cooperating with each other, through structured discussion or in everyday situations" (QUEST). "Staff facilitates positive peer interactions among all children" (ITERS-R).
Positive Regard/Warmth: Positive interactions that are individualized	"Verbal interactions with children are positive" (CDPES). "Provider is warm and nurturing with the children" (APFCCH). "Caregiver shows affection to each child, including gentle touch, kind words, special looks" (QUEST).
Positive Affect: Positive emotional responses by child or caregiver	"Provider expresses positive feelings toward children (laughing and smiling)" (CHELLO). "Children appear to be happy" (APECP). "Focus child was smiling/laughing" (C-COS).
Reciprocity: Multiple responsive exchanges between a caregiver and a child; can be verbal, motori or affective	"Teacher engages children in laughter and smiling through verbal exchanges and/or playful games and activities" (APECP). "Staff have many turn-taking conversations with children (for example, imitate infant sounds in a back-and-forth 'baby conversation'" (ITERS-R). "There is a natural flow in the exchange of information that encourages children to engage in back-and-forth exchanges with the teacher" (CLASS).
Mutuality: Caregiver and child playing/working together	"Caregiver plays interactively with children" (QUEST). "The teacher spends most of her time actively involved with children during free play and planned activities and consistently expands children's involvement. During free play and planned activities, the teacher moves around the room playing with and talking to the children" (CLASS).
Joint Attention: Caregiver and child focusing together on a single object or activity	"In a joint attention episode, both members of a dyad are simultaneously focused on an object or set of objects, while maintaining awareness of the other member's parallel focus" (Markus, Mundy, Morales, Delgado, & Yale, 2000, p. 303). "The amount of time the parent and infant/toddler were looking at/interacting with the same object" (Dodici et al., 2003, p. 127). "Staff engage in educational interaction with . . . individual children" (ECERS-R Revised). "Provider spends quiet, one-on-one time with children" (APECP). "Provider looks at and/or reads book with children daily." "Children are consistently focused on and engaged in free play and planned activities" (CLASS).

Table A3. Definitions and examples of child-caregiver interaction constructs (continued)

Positive or Negative Interaction Construct	
Behavior Regulatory Style/Guidance: Providing behavioral guidelines and prompting desired behaviors; disciplinary styles or parenting styles that help regulate behaviors; the absence of positive behavior guidance may result in overly permissive parenting; in this same construct, negative behavior guidance (such as controlling parenting) may also be measured	"Provider sets clear expectations, and establishes positive, constructive relationships with adults and older children" or "provider sets vague expectations about rules and . . . may use physical action to resolve conflict" (CHELLO). "Directions are positively worded ('Feet belong on the floor'), not just restrictions ('Don't climb on the table')" or "when children misbehave, they are handled abruptly or harshly" (CCIS). "Positive methods of discipline used effectively" or "discipline is either so strict that children are punished or restricted or so lax that there is little order or control" (ITERS-R).
Negative Interaction Constructs	
Detachment: Demonstrating an inability to emotionally connect with one another; disengaged	"Seems distant or detached from the children" (CIS). "Detachment/disengagement" (ORCE). "Predominant focus child/caregiver tone is detached" (CCAT-R).
Intrusiveness: Interrupting the child's activities rather than supporting the child's engagement and exploration of the environment	"The teacher is rigid, inflexible, and controlling in his/her plans and/or rarely 'goes with the flow' of children's ideas; most classroom activities are teacher-driven" (CLASS). "Intrusiveness" (ORCE).
Negative Regard: Negative interactions that are targeted toward another	"Provider's manner may seem harsh or punitive" (CHELLO). "Seems unnecessarily harsh when scolding" (CIS). "Most staff-child interaction is negative" (ECERS-R Revised).
Negative Affect: Negative emotional responses during an interaction	"The teacher consistently displays . . . negative affect" (CLASS). "Predominant focus child tone is upset/crying" (CCAT-R). "Depressive affect" (CCIS).

Source: Halle, Blasberg, Chrisler, and Simkin (2011).

Table A4. Expected correlations between CLASS and simple measures, by subscale

	CLASS								
	1. Positive climate	2. Negative climate	3. Teacher sensitivity	4. Regard for child perspectives	5. Behavior guidance	6. Facilitation of learning and development	7. Quality of feedback	8. Language modeling	Expected no. of correlations
ITERS-R									
1. Space and furnishings									0
2. Personal care routines	+		+	+					3
3. Listening and talking	+		+	+		+	+	+	6
4. Activities	+		+	+		+	+	+	6
5. Interaction	+		+	+	+	+		+	6
6. Program structure				+		+	+	+	4
7. Parents and staff	+				+	+	+		4
CC-IT-HOME									
1. Caregiver responsivity	+		+	+	+	+	+	+	7
2. Acceptance	+	+	+						3
3. Organization				+		+	+	+	4
4. Learning materials				+		+	+	+	4
5. Caregiver involvement	+		+		+		+		4
6. Variety of stimulation				+			+	+	3
MITRCC									
1. Social-emotional	+		+		+			+	4
2. Physical development									0
3. Cognitive			+			+		+	3
Expected no. of correlations for CLASS	9	1	9	9	5	9	9	10	

Notes: Table A4 shows the correlations that are anticipated between the subscales of the instruments, based on the analysis presented in Table 2. "+" refers to a positive and significant correlation between two subscales.

Table A5. Exploratory and Confirmatory Factor Analyses loadings and fit statistics

		Loadings	Model fit	
CLASS	**Domain 1**			
	1. Positive Climate	0.98		
	2. Negative Climate	0.40		
	3. Teacher Sensitivity	0.96	Chi-Square	59.64
	4. Regard for Child Perspectives	0.58	Df	16.00
	5. Behavior Guidance	0.38	RMSEA	0.08
	Domain 2		CFI	0.98
	5. Behavior Guidance	0.53	CD	1.00
	6. Facilitation of Learning and Development	0.92		
	7. Quality of Feedback	0.82		
	8. Language Modeling	0.86		
ITERS-R	**Structural Quality**			
	1. Space and Furnishings	0.69		
	4. Activities	0.64	Chi-Square	46.51
	7. Parents and Staff	0.62	Df	13.00
	Process Quality		RMSEA	0.08
	2. Personal Care Routines	0.61	CFI	0.97
	3. Listening and Talking	0.77	CD	0.95
	5. Interaction	0.87		
	6. Program Structure	0.79		
CC-IT-HOME	**Structural Quality**			
	3. Organization of the Environment	0.48		
	4. Learning Materials	0.61	Chi-Square	33.31
	Process Quality		Df	8.00
	1. Parental Responsivity	0.71	RMSEA	0.09
	2. Acceptance of Child	0.27	CFI	0.94
	5. Parental Involvement	0.66	CD	0.76
	6. Variety in Experience	0.50		

Notes: Factor loadings (standardized estimates) and goodness of fit statistics for two-factor models for CLASS, ITERS-R and CC-IT-HOME. The model for CLASS includes a path from subscale 5 to both domains 1 and 2, and correlated errors between subscales 6 and 8, and 7 and 8. Df refers to the degrees of freedom, RMSEA to the root mean square of approximation, CFI to the comparative fit index, and CD to the coefficient of determination.

Table A6. Ranking correlation matrix for CLASS and simple measures, by subscale

	CLASS								
	1. Positive climate	2. Negative climate	3. Teacher sensitivity	4. Regard for child perspectives	5. Behavior guidance	6. Facilitation of learning and development	7. Quality of feedback	8. Language modeling	Total score
ITERS-R									
1. Space and furnishings	0.18***	-0.02	0.19***	0.10*	0.12**	0.10**	0.06	0.12**	
2. Personal care routines	0.35***	0.10**	0.35***	0.22***	0.33***	0.27***	0.24***	0.27***	
3. Listening and talking	0.42***	0.25***	0.40***	0.24***	0.43***	0.35***	0.30***	0.39***	
4. Activities	0.26***	0.12**	0.23***	0.15***	0.27***	0.22***	0.17***	0.23***	
5. Interaction	0.36***	0.18***	0.34***	0.16***	0.34***	0.26***	0.21***	0.28***	
6. Program structure	0.25***	0.14***	0.27***	0.21***	0.25***	0.20***	0.15***	0.17***	
7. Parents and staff	0.11**	-0.09*	0.09*	0.03	0.08*	0.03	0.06	0.09*	
Total score									0.38***
CC-IT-HOME									
1. Caregiver responsivity	0.32***	0.15***	0.32***	0.12**	0.28***	0.24***	0.21***	0.28***	
2. Acceptance	0.03	0.12**	0.03	-0.05	0.03	-0.01	0.04	0.04	
3. Organization	0.26***	0.12**	0.23***	0.17***	0.26***	0.19***	0.18***	0.19***	
4. Learning materials	0.24***	0.10**	0.24***	0.20***	0.23***	0.21***	0.13***	0.18***	
5. Caregiver involvement	0.24***	0.09*	0.26***	0.14***	0.25***	0.24***	0.21***	0.20***	
6. Variety of stimulation	0.26***	0.07	0.23***	0.17***	0.25***	0.28***	0.22***	0.30***	
Total score									0.38***
MITRCC									
1. Social-emotional	0.32***	0.11**	0.30***	0.16***	0.31***	0.29***	0.24***	0.28***	
2. Physical development	0.14***	0.16***	0.14***	0.06	0.12**	0.08	0.04	0.02	
3. Cognitive	0.27***	0.09*	0.26***	0.19***	0.30***	0.29***	0.20***	0.23***	0.33**
Total score									

Notes: Centers were ranked from worst to best according to their score. This table shows the Pearson correlation coefficients between the ranking of the centers for each subscale. Coefficients significant at * p<0.10, ** p<0.05, and *** p<0.01.