Author(s): Pakarinen, Eija; Malmberg, Lars-Erik; Poikkeus, Anna-Maija; Siekkinen, Martti; Lerkkanen, Marja-Kristiina

Title: Investigating applicability of ratings of indicators of the CLASS Pre-K instrument

Year: 2023

Version: Accepted version (Final draft)

Please cite the original version:

Pakarinen, E., Malmberg, L.-E., Poikkeus, A.-M., Siekkinen, M., & Lerkkanen, M.-K. (2023). Investigating applicability of ratings of indicators of the CLASS Pre-K instrument. International Journal of Research and Method in Education, 46(3), 231-247. https://doi.org/10.1080/1743727x.2022.2128741

Investigating Applicability of Ratings of Indicators of the CLASS Pre-K Instrument

Eija Pakarinen[1], Lars-Erik Malmberg[2], Anna-Maija Poikkeus[1], Martti Siekkinen[3], & Marja-Kristiina Lerkkanen[1]

[1]Department of Teacher Education, P.O. Box 35, 40014 University of Jyväskylä, Finland [2]Department of Education, University of Oxford, 15 Norham Gardens, Oxford, OX2 6PY, UK

[3]Philosophical Faculty, School of Applied Educational Sciences and Teacher Education, P.O. Box 111, 80101 Joensuu, University of Eastern Finland, Finland

eija.k.pakarinen@jyu.fi; lars-erik.malmberg@education.ox.ac.uk; anna-maija.poikkeus@jyu.fi; martti.siekkinen@uef.fi; marja-kristiina.lerkkanen@jyu.fi

Acknowledgements

This study was financed by grants from Academy of Finland (Nr. 277299 for 2015–2017, Nr. 292466 for 2015–2019; Nr. 268586 2013–2017).

For any correspondence, contact: Eija Pakarinen, Department of Teacher Education, University of Jyväskylä, P.O. Box 35, 40014 University of Jyväskylä, Finland. E-mail: eija.k.pakarinen@jyu.fi; phone: +358-40-8053520.

Investigating applicability of ratings of indicators of the CLASS Pre-K instrument

When classroom observations are increasingly used for accountability and evaluation purposes, a deeper understanding of the psychometric properties of such measurement tools is needed. The present study took a unique approach to examine the psychometric properties of a commonly used classroom observation measure by testing the reliability of indicators for higher-order constructs (i.e., dimensions). We investigated the reliability of indicator ratings of the Classroom Assessment Scoring System (CLASS) Pre-K instrument in Finnish kindergarten and first grade classrooms. Twenty-one observer pairs rated 838 segments identified from the 413 lessons of 48 teachers. Variance components models were specified to investigate variance proportions of each indicator and dimension. The results showed that most observer disagreement was found for the instructional support domain. Observers disagreed relatively more depending on the teacher they observed. There is a clear need for additional understanding on how observers process information on the complex elements of classroom interaction in order to improve training programs and the reliability and accuracy of the assessment procedure.

Keywords: Classroom Assessment Scoring System Pre-K; indicators; inter-rater reliability; variance components models

Introduction

Observations of process quality have increasingly been applied to study teacher–child interactions in early childhood education and primary school settings and more widely used to improve teacher effectiveness. Results stemming from standardized instruments enable us to differentiate between schools, classrooms and teachers and can be used for both high-stakes (e.g., teacher recruitment and bonuses) and low-stakes (e.g., teacher professional development) decisions (Curby et al. 2016). As assessment of the quality of teacher–child interactions is only as good as the quality of the indicators used (e.g., validity of measures and degree of needed inference), observer reliability (e.g., training, inter-rater agreement and consistency over time) and the accessibility of the contexts of interest (e.g., peer interactions and small group discussions), there is a clear need to deeply understand the decision-making processes during observation as well as psychometric properties of the measurement tools.

The Classroom Assessment Scoring System (CLASS) (Pianta, La Paro and Hamre 2008), which is most widely used in the United States (e.g., Hamre et al. 2013) and in many other countries outside the US (e.g., Cadima, Leal and Burchinal 2010; Hu et al. 2016; Leyva et al. 2015; Suchodoletz et al. 2014), has shown good psychometric qualities at the level of domains and dimensions (e.g., factor structure and internal consistency) (for a meta-analysis, see Li, Liu and Hunter 2020). Observers using CLASS are trained to a high standard (i.e., high inter-rater agreement, intensive training and annual recertification of observers) (Cash et al. 2012; Pianta and Hamre 2009; Pianta, La Paro and Hamre 2008). During the training, observers are trained to detect various behavioural indicators (e.g., children's smiles or shared activities) and to disregard their pre-conceptions concerning optimal teaching or structuring of learning situations (e.g., "What is my favourite teacher like?"). Observers are trained to take

notes on the indicators, which help them to distinguish between a low-, average- or high-quality score for each observed dimension.

While there is ample information on the domain and dimension levels, there are a lack of studies investigating the validity and reliability of the indicators on which decisions of dimension scores are based. Jensen et al. (2020), for example, concluded that a major limitation of the CLASS instrument is that observers assign scores at the dimension rather than the indicator level. In the present study, we observed 48 classrooms in order to investigate whether indicator ratings could reliably predict dimension scores of CLASS. This analytic approach extends previous studies on the measurement quality of CLASS by providing information needed to understand the psychometric properties of the measurement and to develop observer training.

Classroom quality and the CLASS instrument

The Teaching Through Interactions (TTI) framework (Hamre et al. 2013) conceptualizes effective teacher–child interactions along three broad domains: emotional support, classroom organization and instructional support which can be measured by CLASS (Pianta, La Paro and Hamre 2008). Emotional support refers to a positive tone to interactions and a warm and supportive classroom climate. Emotionally supportive teachers are sensitive and responsive to children's needs and provide children with appropriate levels of leadership and autonomy (Pianta, La Paro and Hamre 2008). Classroom organization refers to teachers' effective management of time and attention and to setting clear rules and routines (Yates and Yates 1990). In addition to providing a structure for learning, teachers with high classroom organization skills also promote students' motivation and provide inherently interesting activities for the children (Pianta, La Paro and Hamre 2008). Instructional support captures the quality of feedback, stimulation of thinking skills and reasoning in the classroom as well as

explicit linking of content knowledge with meaningful contexts (Pianta, La Paro and Hamre 2008). Some versions of the CLASS instrument include a fourth domain, student engagement, which refers to the degree to which students are actively engaged in learning tasks.

Increasing evidence shows the importance of the quality of teacher–child interactions for child outcomes, providing evidence of the predictive validity of CLASS. Emotional support is linked to gains in pre-schoolers' academic skills (Burchinal et al. 2010), social competence (Curby, Rimm-Kaufman and Ponitz 2009; Mashburn et al. 2008) and behavioural regulation (Merritt et al. 2012). Classroom organization is related to pre-schoolers' task orientation (Dobbs-Oates et al. 2011) and first graders' print awareness, vocabulary (Cadima, Leal and Burchinal 2010) and literacy gains (Ponitz et al. 2009). Furthermore, high instructional support has been linked to pre-schoolers' pre-literacy and pre-math skills (Mashburn et al. 2008) and their word reading progress (Curby, Rimm-Kaufman and Ponitz 2009).

The different versions of the CLASS instrument all measure the three domains of teacher–student interactions, but they somewhat differ in the specific dimensions within each domain and the corresponding indicators (see Table 1). CLASS Pre-K (Pianta, La Paro and Hamre 2008) is a well-validated observation instrument for children in early education and preschool settings. The published measure identifies 10 dimensions of teacher–child interactions. Dimensions are comprised of four to five indicators, which describe the concrete behavioural markers, that is, descriptive anchors for each dimension. Typically, researchers score at the dimension level, which is commensurate with the CLASS training and manual. Although global codes tend to be better predictors of child outcomes, they are often more difficult to learn as an observer and perhaps more prone to reliability issues given the higher level of inference needed

to apply the codes. Understanding the indicators (i.e., what a coder is expected to use when making inferences) might be helpful for improving the properties of a global, integrative, oriented measure. However, for those interested in understanding psychometric properties of the instrument and teacher professional development, the dimensions may prove to be too broad of a construct for observer training and specific feedback. Therefore, assessment at the indicator level may provide a useful platform for discussions in training and detecting potential differences between observers. Unpacking CLASS at the indicator level allows for testing implicit assumptions of CLASS (i.e., that these behavioural indicators reflect higher-order dimension constructs) and identifying indicators that may need more or less attention when training observers to high standards of reliability.

Structural validity of the CLASS instrument

The growing popularity of the TTI framework in research and practice has led to numerous studies of the psychometric properties of the CLASS instrument (see Li, Liu and Hunter 2020). The literature has provided strong evidence of the structural validity of CLASS at the levels of domains and dimensions. Studies have also indicated high reliability in terms of internal consistencies (Cronbach's alphas) for the CLASS domain and dimension scores (e.g., Hamre et al. 2013; Hu et al. 2016; Pianta, La Paro and Hamre 2008; Suchodoletz et al. 2014). Hamre et al. (2013) showed that the theoretical three-factor model (consisting of emotional, organizational and instructional support) provided the best fit for the data of over 4,000 early childhood and elementary classrooms. Similarly, Sandilos and DiPerna (2014) demonstrated among 417 kindergarten classrooms that a three-factor structure for the CLASS K–3 instrument provided the best fit after some modifications to the original CLASS model. Furthermore, Downer et al. (2012) reported that the three-factor structure of the CLASS

instrument applied equally well across classrooms with different Latino and dual language learner compositions.

Some studies that have examined the construct validity of the CLASS instrument outside the US indicated that CLASS is also applicable in other cultural and educational contexts, at least with some modifications. For example, Leyva et al. (2015) reported three distinctive but interrelated domains of teacher–child interactions among 91 Chilean preschool classrooms. Suchodoletz et al. (2014) and Stuck, Kammermeyer and Roux (2016) demonstrated that the same three-factor structure provided the best fit for data of German preschool classrooms. Hu et al. (2016) showed that the three-factor model fit the data well in a sample of 118 Chinese kindergarten classrooms. However, Jensen et al. (2020) demonstrated with mixed methods that a revised three-factor model fit the data better than the original theoretical model in a sample of Mexican K–1 classrooms.

Previous research on the structural validity of CLASS has, however, some limitations. First, because CLASS is now being widely used internationally, information on the construct validity of the CLASS instrument in different educational and cultural contexts is needed to confirm the applicability of the instrument. Second, although evidence supports the three-factor structure across different contexts, the model's fit has not been ideal, and some modifications to the model have been needed (Leyva et al. 2015; Malmberg et al. 2010; Suchodoletz et al. 2014). Unpacking CLASS at the indicator level might help reveal the reasons for slight modifications in factor analytic work with the CLASS instrument across a variety of different countries. Third, most of the previous studies have ignored the hierarchical structure of the data by using only teacher-level aggregated scores. Moreover, no studies thus far have investigated observers' rating processes from observations to dimension ratings, that is, on which

indicators they base their scores. Although observers are trained to a high standard, there is a lack of information on how best to train observers to use the indicators to assess the quality of interactions (e.g., information on variation between observers across segments). During the CLASS training, observers are instructed to pay attention to focal behavioural incidents and to take notes as justification for distinguishing between low-, average- or high-quality scores for each dimension on the basis of these indicators. At least to our knowledge, this study is one of the first attempts to investigate the reliability at the level of indicators. The current study provides unique data based on two observers' scores for each observed lesson at the indicator level, which allows for investigation of the reliability of indicators. The present analytic approach allows us to test implicit assumptions that the indicators reflect higher-order dimensions and to identify indicators that may need more or less attention in training observers to reliability standards.

Sources of variation in CLASS scores

The observed score representing the quality of teacher–child interactions has a complex variance structure comprising multiple sources of variability related to teachers and classrooms, days within classrooms, occasions within days (lessons, segments), observers and variances due to the interactions of these (Mashburn, Downer et al. 2014; McCaffrey et al. 2015). For example, the content and activities of a specific lesson may produce variability in ratings (Bejar 2012). In addition, the events from or impressions left by the prior segments are likely to affect the scores of a subsequent lesson (Mashburn, Meyer et al. 2014). Although observers using the CLASS instrument are trained to a high standard, there are some potential sources of error in their scores, such as observer bias, calibration and observer leniency. Cash et al. (2012) showed that the majority of variation in observer calibration takes place at the observer level. For

example, an observer may have difficulty understanding and applying the coding protocol (Bejar 2012). Recent work by Bell et al. (2014) demonstrated that observers have the highest agreement and accuracy when scoring dimensions belonging to the classroom organization domain and the least accuracy on the instructional support and emotional aspects of interactions domains, suggesting that dimensions (which require lower-inference judgments) are easier to score than domains.

Observers may also differ in their overall severity or leniency in their ratings (Styck et al. 2021). Previous work on observation protocols has repeatedly found observer errors to be a major source of the variability in scores (Casabianca et al. 2015; Casabianca et al. 2013; Mashburn, Downer et al. 2014). In previous studies, observers tended to agree more with one another on organizational aspects of teacher–student interactions (Bell et al. 2012), whereas instructional dimensions were the most challenging to rate reliably and accurately (Bell et al. 2014; Gitomer et al. 2014). Observers may also differ in their beliefs and opinions of what counts as high-quality teaching as well as in their own experiences with teaching (Bejar 2012).

Typically, 15–20% of the CLASS data are double coded for investigating inter-rater reliability using adjacent agreement (ratings within 1 point on the 7-point rating scale; Pianta, La Paro and Hamre 2008), exact agreement and intra-class correlation coefficients (ICCs). Inter-rater reliability (adjacent agreement) for CLASS dimensions typically ranges from .72 to .89 (Brown et al. 2010). According to the CLASS manual (Pianta, La Paro and Hamre 2008), 80% agreement within 1 point is acceptable. Styck et al. (2021) indicated that inter-observer agreement calculated as exact agreement at the domain level at Grade 1 and 2 was 65.6% for emotional support, 53.1% for classroom organization and 38.0% for instructional support. The average exact agreement across the cycles ranged from 35% to 40% (M = 37%) in a kindergarten study by

Mantzicopoulos et al. (2018). Inter-rater reliability in terms of intra-class correlations, in turn, has been shown to vary from .64 to .87 at the level of dimensions and from .76 to .87 at the level of domains (Hamre et al. 2014). Hu et al. (2016) demonstrated that ICCs at the dimension level, indicating inter-rater reliability, varied from .82 to .91.

Recent studies have also investigated inter-rater reliability of the CLASS instrument in terms of variance components models. In a sample of kindergarten classrooms, Mantzicopoulos et al. (2018) identified a teacher by observer interaction that accounted for 5.4% to 15.9% of variance across domain scores. Furthermore, Jensen et al. (2020) demonstrated substantial teacher by observer interaction effects, which were highest for emotional support dimensions and lowest for instructional support dimensions. In a study by Mashburn, Downer et al. (2014), ratings varied considerably across observers, and this between-observer variance was substantial for instructional support (18%) and emotional support (14%) but relatively small for classroom organization (4%). The greatest source of variance for the CLASS domains (between 27% and 33% of the total variance) was observer by occasion variance.

Although there is some research on the possible sources of variability in CLASS scores, we are far from understanding the coding processes of observers, that is, to what extent observers differ at the level of indicators. So far, observer bias and variability in scores has only been studied regarding the CLASS Secondary version. Systematic research informing about these issues using the CLASS Pre-K instrument is missing. In the present study, we investigated the reliability of indicator ratings of the CLASS Pre-K instrument. Unpacking CLASS at the indicator level enables us to test implicit assumptions that these indicators reflect higher order dimension constructs and to identify indicators that may need more or less attention when training observers to reliability standards.

The present study

The present study investigated the reliability of indicators of the CLASS instrument by determining how much of the variance is attributable to observers. Following generalizability theory (Cronbach et al. 1972; see also Jensen et al. 2020; Mantzicopoulos et al. 2018; Mashburn, Downer et al. 2014), we specified variance components models for three reasons. First, the quality of the observation scores affects the generalizability of findings. Unreliability of observation measures (e.g., observed disagreement) can increase Type II errors, that is, the effects of observed quality on student outcomes is diminished. Second, the magnitudes of the variance proportions can then be used for making informed decisions about how a research methodology can be implemented in subsequent studies (e.g., how many observer pairs are needed, how observation training can be carried out and which indicators require further emphasis when training observers). Third, when partitioning variance by its main sources, it is preferable for the variance between teachers to be the largest and the observer, observer by episode and teacher by observer variances to be smaller. As this was, to our knowledge, one of the first studies to investigate indicator-level reliability, we proposed the following research questions without specific hypotheses:

(1) How much variance do observers account for in the scoring of each indicator?
(2) How much variance do observers account for in the scoring of each dimension?

Method

Participants and procedures

This study used observer inter-rater reliability data of CLASS assessments carried out in kindergarten and first grade classrooms. The data were drawn from a larger follow-up study. Forty-nine kindergarten teachers (47 female, 2 male) with 515 children as well as

16 first grade teachers (15 female, 1 male) with 258 students participated in the classroom observations (a total of 65 classrooms; see Table 2 for demographics).

Two observers per teacher observed five to six live lessons (divided into 20-minute segments, i.e., coding cycles) during two consecutive days: two observers by 10 segments by 65 teachers, for a total of 980 possible observed segments. Observers were assigned to classrooms primarily based on observer availability. Observers (29 in total) were instructed to record their ratings of indicators on the scoring sheet. We included ratings of segments that had at least 19 (out of the maximum 56) indicator ratings available, giving an overall rate of missingness of 4.6% in the sample: 838 segments during 413 lessons of 65 teachers with 20 observer pairs. The data fulfilling these criteria consisted of the lessons of 48 teachers. All observed segments had data from both observers available, and there were between three and 10 segments observed per teacher (M = 8.73; SD = 1.89).

Measures

We used an unpublished version of the CLASS instrument, which included 11 dimensions of observed teacher–child interactions (10 dimensions in the current version; see Table 1). Each of them had several indicators (total of 56 indicators; 42 in the current version). Although in the current version there are typically fewer indicators describing dimensions, the content of the dimensions is the same. Previous indicators have not been dropped from the published version of the manual, but some indicators have been merged. The only exception is quality of feedback, for which there are more indicators in the current version of the manual.

Each dimension was rated on a 7-point scale (1–2 = low; 3–5 = mid; 6–7 = high). The reliability training is geared towards reaching habitual recognition of behavioural expressions that are coded as indicators of a dimension during the

observation phase of a segment. During the subsequent coding phase, as instructed in the manual, the observer weighs the frequency, duration and intensity of each indicator of each dimension. In the present study, the observers were asked to also assign ratings on a scale of 1 to 7 for each of the indicators before assigning the final ratings for the dimensions. This produced two-tiered ratings data at the level of indicators and their respective dimensions. The internal consistencies (alphas) of the a priori constructs based on their dimensions were .72 for emotional support (positive climate, negative climate [reverse coded], teacher sensitivity and regard for student perspectives), .68 for classroom organization (behaviour management, productivity and instructional learning formats) and .83 for instructional support (concept development, quality of feedback and language modelling).

Procedures

Before starting the observations, 29 observers were carefully trained. Ratings that were within 1 point of each other were considered to reflect an acceptable degree of accuracy (Pianta, La Paro and Hamre 2008). In cases where the pairs of observers had discrepancies greater than 1 scale point between their codings, extra coding practice in authentic classrooms was required, and a meeting was arranged to monitor the inter-rater agreement.

Observations were completed in 30-minute cycles. A pair of trained observers first observed a 20-minute period while making notes on indicators on a separate sheet of paper, and in the subsequent 10-minute period, they recorded both their codings of indicators and the dimensions on the scoring sheet independently of each other. Inter-rater reliabilities (intraclass correlations) between the observers at the level of dimensions varied from .63 to .96 in kindergarten and from .40 to .96 in the first grade.

Analysis strategy

We specified a series of variance components models in SPSS Statistics 21. We partitioned the variance into main effects of teacher (*t*), observer (*o*) and episode (*e*) and interaction effects of teacher by observer, teacher by episode and observer by episode. The three-way interaction of teacher by observer by episode was the implicit residual. Following Brennan (2011) and Mashburn, Downer et al. (2014), we estimated the variance components of the fixed effect model as:

$$y_{toe} = \mu + v_t + v_o + v_e + v_{to} + v_{te} + v_{oe} + v_{toe} \quad (1)$$

The variances were estimated as:

$$\sigma^2(y_{toe}) = \sigma^2(v_t) + \sigma^2(v_o) + \sigma^2(v_e) + \sigma^2(v_{to}) + \sigma^2(v_{te}) + \sigma^2(v_{oe}) + \sigma^2(v_{toe}) \quad (2)$$

We examined the variance proportions of each indicator and dimension score in turn. Our data were hierarchically organized with ratings from both observers (n_{sti} = 838 rated segments) nested within each lesson/episode (n_{ti} = 413 observed episodes) nested within teachers (n_i = 48 teachers). For comparison purposes, we also calculated observer reliability in terms of exact agreement and weighted Kappas for the indicators (Tables 3–5).

Results

We first observed the proportion of variance of the observers at the level of indicators. On average, the most variance was between teachers (42.5%) and episodes of teachers (i.e., the T × E component; 38.9%) indicating individual differences between teachers but also showing that teachers varied from one episode to another in their quality of teacher–child interactions. There was relatively little variance between episodes (1.6%).

As our focus is on the observer variance, we noted that there was an average variance component of 8.6% (from 0 to 32.3%) and median variance proportion of 6.9% across all the items. Observers were consistent from one episode to another (i.e., the O × E component; average: 0.8%, range: 0 to 7.0%; *Md* = 0%). Observers disagreed relatively more depending on the teacher they observed (i.e., the T × O component; average: 7.7%, range: 0 to 40.0%; *Md* = 0.070). It is important to scrutinize the indicators with a relatively high proportion of disagreement. Given the skewed distribution of the variance components, we used the median variance proportion (*Md* = 0.069) of observer component as a cut-off for the interpretation of a relatively high proportion of observer disagreement. To scrutinize the indicators having greater-than-desirable observer disagreement, we inspected values outside the third quartile plus 1.5 times the interquartile range for the observer, O × E, and T × O variance components. The cut-offs were 0.235 for observer, 0.028 for observer by episode and 0.168 for observer by teacher components. These values with greater-than-desirable variability attributable to observers are shown in bold in Tables 3–5.

When we inspected the items within the separate dimensions, the following average observer variance components emerged. Observers accounted on average for 6.5% (Min = 0.7%, Max = 14.0%) of the variance in the 22 emotional support indicators, 6.3% (Min = 0%, Max = 29.8%) in 17 classroom organization indicators, 14.9% (Min = 7.3%, Max = 32.3%) in 19 instructional support indicators, and 19.8% (Min = 0%, Max = 32.3%) in two student engagement indicators. The following indicators of emotional support (Table 3) were above the median observer variance component indicative of disagreement: one indicator from the positive climate dimension (peer interaction), three indicators from the negative climate dimension (sarcasm and disrespect, negativity not contained to events and severe negativity), two

indicators from the teacher sensitivity dimension (responsive and addresses problems) and two indicators from the regard for student perspectives dimension (support of autonomy and restriction of movement). In addition, peer interaction, punitive control and negativity not constrained to events had particularly high observer by episode variance components. Escalating negativity and severe negativity had particularly high teacher by observer variance components.

There were two indicators in the classroom organization domain (Table 4) that had an above median observer variance component: the transitions and managerial tasks indicators from the productivity dimension and the modalities indicator from the instructional learning formats dimension. In addition, managerial tasks had particularly high observer variance components. Proactivity and loss of time had high observer by episode variance components, and clear behavioural expectations had particularly high teacher by observer variance components. In addition, the sustained engagement indicator of the student engagement dimension had an above median observer variance proportion (Table 5).

Most observer disagreement was found for the instructional support domain (Table 5). All concept development dimension indicators, quality of feedback indicators and language modelling indicators showed above median observer variance component. Closer inspection revealed that feedback loops and self and parallel talk indicators had particularly high observer variance components, and specific feedback had a particularly high observer by episode variance component.

In addition, average exact agreement and weighted Kappa coefficients are reported in Tables 3–5. The average exact agreement ranged from .61 to .89 (M = .74). Two indicators (self and parallel talk as well as repetition and extension) of the language modelling dimension had the lowest exact agreement scores in our sample.

The average exact agreement by observer pairs was 83%. Most of the weighted Kappas ranged from .27 to .52 (M = .37), which can be interpreted as fair to moderate (Cohen, 1968; McHugh, 2012) (with the exception of severe negativity = .081). Several indicators of the negative climate dimension had the lowest Kappa coefficients, demonstrating fair inter-rater reliability. Furthermore, simple inter-rater reliability in terms of the ICC was excellent (.97).

Discussion

Information on the reliability of carrying out observational ratings in classrooms is clearly needed, as observational measures are increasingly being used to describe and evaluate teacher effectiveness. The aim of this study was to investigate the reliability of indicator ratings of CLASS scores by determining how much variability is attributable to observers. Thus, we go beyond previous studies of domain- and dimension-level analysis of measurement qualities of the CLASS instrument. Following Generalizability Theory, the results of variance components models indicated that most of the indicators of the CLASS dimensions were reliable, showing that relatively little variability was attributable to observers. However, there were some dimensions that had greater-than-desirable variability attributable to observers and could be improved in terms of observer agreement. There were also some differences in the reliability of indicators at different levels (i.e., segment, episode and teacher), indicating variability between observers. The results of the present study are of particular importance as this is the first study to provide information on ratings of indicators as complementing the dimension ratings, as suggested by Jensen et al. (2020).

This study examined what proportion of the variance observers account for in scoring each indicator. Observers accounted for approximately 7% of the variance in the emotional support, 6% in the classroom organization, 15% in the instructional support

and 20% in the student engagement indicators. Regarding emotional support, one indicator of the positive climate, three indicators of the negative climate, two indicators of the teacher sensitivity and two indicators of the regard for student perspectives dimensions were above the median observer variance component indicative of disagreement. The weighted Kappa coefficients were also lowest for the negative climate indicators, indicating fair inter-rater reliability. Most observer disagreement was found for the instructional support domain, as all indicators showed above median observer variance components. The average exact agreement scores were aligned and provided evidence on the triangulation of various reliability estimates; the average exact agreement scores were lowest for indicators belonging to the instructional support domain. Jensen et al. (2020) also indicated that observer error was particularly large for dimensions of instructional support.

With a more stringent inspection, three indicators showed greater-than-desirable variability attributable to observers: managerial tasks (indicative of productivity), feedback loops (indicative of feedback) and self and parallel talk (indicative of language modelling). Jensen et al. (2020) reported that at the dimension level, observers varied the most in concept development, quality of feedback and language modelling. In addition, six indicators demonstrated greater-than-desirable observer by episode interaction: peer interaction, punitive control, negativity not contained to events, proactivity, loss of time and specific feedback. This result is indicative of between-observer differences in ratings of the indicators varying from one episode to the next. Furthermore, negativity escalates, severe negativity and clear behavioural expectations had particularly high teacher by observer interaction, suggesting that the between-teacher differences in these indicator ratings varied between observers. Jensen et al.

(2020) showed that the teacher by observer interaction effects were highest for the emotional support dimensions and lowest for the instructional support dimensions.

Reliability problems at the indicator level emerged in all three CLASS domains, but to a lesser extent in the domain of classroom organization, implying that for this domain, behavioural markers are least likely to be based on inferences. These results align with the previous literature showing that the dimensions of classroom organization are easy to observe, as they require lower-inference decision making (Bell et al. 2014). In a similar vein, Hu et al. (2016) showed that the dimensions belonging to the classroom organization factor were highly reliable indicators of the latent domains in a Chinese kindergarten sample. This may also be because children typically behave well at this stage, and thus teachers do not need to apply their classroom organization skills.

There are some possible explanations for these less optimal results. First, the dimension measuring negative climate has been found to be somewhat problematic in previous studies as well. For example, it has shown low factor loading and low correlations with the emotional support domain both in German (Suchodoletz et al. 2014) and Chilean (Leyva et al. 2015) preschool samples. In fact, in recent versions of the CLASS Secondary instrument, negative climate was moved to the domain of classroom organization (Pianta et al. 2012). The lower psychometric properties concerning the negative climate dimension suggest that clarification of the coding scheme is especially pertinent when the dimension taps practices that may vary by cultural contexts, norms of communication and expression of affects. Our results might suggest that the behavioural markers of negative climate may be culturally sensitive and less easily transferable from one context to another. It may also be that sarcasm/disrespect and severe negativity are not present in classroom situations, as previous studies have indicated that there is low variability in negative climate (Leyva

et al. 2015). Typically, individual children may express some negative affect, but negativity does not escalate.

In addition, the dimensions measuring instructional support, for which the proportion of variance due to observers was not optimal, may be difficult to rate reliably, especially if the observers have different background knowledge of children's conceptual development. The observers may, for instance, vary in terms of their views of what they construe as a 'concept' in the early childhood education context, where instruction does not necessarily involve accuracy of the content or clear but abstract concepts, as in the later years. Moreover, indicators of instructional support dimensions may be less evident in observational situations than, for example, indicators related to emotional support dimensions. Consequently, the results suggest the need to focus on the less evident dimensions and their behavioural markers when training observers and monitoring the reliability of their observations. Perhaps more detailed and explicit descriptions, examples and clarification of the coding criteria are needed regarding the higher-inference dimensions. In a similar vein, previous studies have demonstrated that observers tend to agree more with one another on organizational aspects of teacher–student interactions (Bell et al. 2012), whereas instructional dimensions have been the most challenging to rate reliably and accurately (Bell et al. 2014; Gitomer et al. 2014).

Quality of feedback may also vary between play-centred activities and more academically oriented activities. In a related finding, McCaffrey et al. (2015) demonstrated that errors in CLASS Secondary scores made by two observers on the same lesson had a factor structure that was different from the factor structure at the teacher level. These findings suggest that there is a clear need to investigate the structural validity of CLASS scores at the different levels of data. Connection to the real world and feedback loops are the indicators that require high inference. Another

explanation for the less optimal model fits of these indicators could be that these indicators are not typically present in observed classroom situations. Also, previous studies have demonstrated that scores for instructional support are not optimal but typically in the low to mid-range (Hamre et al. 2013; Hu et al. 2016; Leyva et al. 2015; Suchodoletz et al. 2014).

The present study also examined what proportion of the variance observers accounted for in the scoring of each dimension. On average, the most variance was found between teachers and for episode by teacher interactions, indicating individual differences between teachers but also showing that the quality of teacher–child interactions varies from one episode to another (Malmberg et al. 2010). There was relatively little variance between episodes. The results showed that observers were consistent from one segment to another but disagreed relatively more depending on the teacher they observed. Similarly, Mantzicopoulos et al. (2018) identified a teacher by observer interaction that accounted for 5.4% to 15.9% of variance across CLASS domains.

The results indicated that, on average, observers accounted for the most variance in dimensions of the instructional support domain. Mashburn, Downer et al. (2014) showed that between-rater variance was substantial for the instructional and emotional support dimensions and relatively small for the classroom organization dimension. It has been suggested that the instructional support dimension requires higher-inference decisions (Bell et al. 2014). For example, Cash et al. (2012) suggested that observers who are highly experienced classroom teachers may unknowingly be influenced by their own ideas regarding instruction or classroom management.

The results may indicate that coding of individual lesson segments (i.e., cycles) includes more situational variation that is not related to teachers' general pattern of

teaching (see McCaffrey et al. 2015). Similarly, McCaffrey et al. (2015) suggested that the factor structure may be different at different levels of the data. In other words, the sources of variation have a hierarchical structure, with segments nested within lessons, lessons nested within classrooms and classrooms nested within teachers. In addition, Gitomer et al. (2014) indicated that the variance is greater for lesson-level scores than for classroom-level scores, and observer agreement rates are slightly higher at the lesson level than at the segment level. Thus, more research is needed on what evidence observers use to make their judgements.

Overall, the results suggest that observers may weigh different indicators of the quality of teacher–child interactions in their ratings. It seems that different levels of data (segment, episode and teacher) contain different information, and future studies should take these levels into account, as suggested by Jensen et al. (2020). Therefore, more studies are needed to investigate the complex nature of observer ratings. Discussion among observers at the training phase is valuable for helping them to become more conscious of the 'salient' coding criteria that they apply, that is, how they process the information of observations of behavioural markers to assign their codings. In the present study, there were some differences in the applicability of indicators at different levels (segment, episode and teacher). To gain additional understanding of the process of utilizing indicator ratings, stimulated-recall interviews or a thinking aloud method could be applied (Bell et al. 2014).

Observations of teaching are commonly used for teacher evaluation in the United States. It is critical to ensure that the observational ratings are accurate and reliable, particularly in high-stakes environments, such as employment decisions. The evidence shows that observer reliability remains a persistent problem when rating instruction using classroom observations (Casabianca, Lockwood and McCaffrey 2015;

Casabianca et al. 2013; Hill, Charalambous, and Kraft 2012). Therefore, emphasis should be on developing observer trainings and coding protocols.

These results have some practical implications. First, some CLASS dimensions may require more observers and observation segments than others to achieve adequate reliability. Practical constraints, however, may limit the possibility of how many days, lessons and ratings per lesson can occur. Second, more observer training time may be needed to identify and calibrate observations related to instructional support. By recording the indicator ratings, the possible systematic biases and differences in interpretation between observers at the level of behavioural markers in the training phase can easily be detected. Focusing on the indicator ratings might help to improve observer reliability. Reviewing and discussing examples that the trainers introduce can clear up any misguided conceptions that may result from inexperience with the educational setting or the concepts used to describe the criteria. Training typically involves reviewing the scoring rubrics, giving examples, providing opportunities to practice scoring followed by discussion and feedback as well as assessing calibration to scores assigned by expert observers. Third, our finding that indicators differed in their reliability at different levels (i.e., segments and episodes) indicates that there is a need for concrete examples of activities when training observers to elucidate subtle variations in interactional quality. Classroom situations contain unanticipated elements, and even when observers have learned coding criteria by rote, unexpected events may bring uncertainty when assigning codes according to rubrics. Spending more time during observer training on the indicators that showed higher than median variance components might help to mitigate the influence of observers' previous experiences and opinions and to improve observer reliability.

Limitations

The present study has some limitations that need to be considered. First, we used a previous version of the CLASS instrument, which included more indicators than the current version. Although in the current version there are typically less indicators describing each dimension, the content of the dimensions is the same. Indicators have not been dropped from the published version of the manual, but some indicators have been merged. Second, the sample size was small. The original sample was reduced to 48 teachers, as not all of the observers provided sub-indicator ratings. Third, the CLASS Pre-K instrument was used in the assessments, although the sample also contained first grade classrooms. The decision to use the CLASS Pre-K measure was, however, based on the fact that the structure and indicators in the CLASS K–3 version are identical. Fourth, it should be noted that due to the observations being live, the same observers rated all the lessons of a certain teacher. This could have an impact on the estimates; for instance, general impression halo error could have played a role and decreased observer by episode interaction. A perfect design would be a random assignment of observers to episodes. However, for practical reasons, this is rarely the case. Fifth, additional sources of variation between the observers could include, for example, their educational background, knowledge and familiarization with early childhood education (e.g., extent of teaching experience). These should be accounted for in future studies.

Conclusions

These findings are of particular importance as they represent the first evidence on the applicability of ratings of indicators of the CLASS instrument. There is an evident need to understand and improve the psychometric properties of the CLASS measure. Overall, this study supports the reliability of indicator ratings. However, the results also suggest

that there are some issues related to specific indicators that should be taken into account when training observers and applying the CLASS instrument, particularly in other cultural and educational contexts. There is a clear need to gain additional understanding of how observers process information on the complex elements of classroom interaction in order to improve training programs and the reliability and accuracy of the assessment procedure.

References

Bejar, I. I. 2012. "Rater Cognition: Implications for Validity." *Journal of Educational Measurement* 31(3): 2–9. doi: 10.1111/j.1745-3992.2012.00238.x

Bell, C. A., D. H. Gitomer, D. H. McCaffrey, B. K. Hamre, R. C. Pianta, and Y. Qi. 2012. "An Argument Approach to Observation Protocol Validity." *Educational Assessment* 17(2-3): 62–87. doi: 10.1080/10627197.2012.715014

Bell, C. A., Y. Qi, A. Croft, D. Leusner, D. McCaffrey, D. H. Gitomer, and R. C. Pianta. 2014. "Improving Observational Score Quality: Challenges in Observer Thinking." In *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project*, edited by K. Kerr, R. Pianta, and T. Kane, 50–97. San Francisco, CA: Jossey-Bass.

Brennan, R. L. 2011. "Generalizability Theory and Classical Test Theory." *Applied Measurement in Education* 24(1): 1–21. doi: 10.1080/08957347.2011.532417

Brown, J. L., S. M. Jones, M. LaRusso, and J. L. Aber. 2010. "Improving Classroom Quality: Teacher Influences and Experimental Impacts of the 4Rs Program." *Journal of Educational Psychology* 102(1): 153–167. doi: 10.1037/a0018160

Burchinal, M., N. Vandergrift, R. C. Pianta, and A. J. Mashburn. 2010. "Threshold Analysis of Associations between Child Care Quality and Child Outcomes for Low-Income Children in Pre-Kindergarten Programs." *Early Childhood Research Quarterly* 25: 166–176. doi: 10.1016/j.ecresq.2009.10.004

Cadima, J., T. Leal, and M. Burchinal. 2010. "The Quality of Teacher–Student Interactions: Associations with First Graders' Academic and Behavioral Outcomes." *Journal of School Psychology* 48(6): 457–482. doi:10.1016/j.jsp.2010.09.001

Casabianca, J. M., J. R. Lockwood, and D. F. McCaffrey. 2015. "Trends in Classroom Observation Scores." *Educational and Psychological Measurement* 75(2): 311–337. doi: 10.1177/0013164414539163

Casabianca, J. M., D. F. McCaffrey, D. Gitomer, C. Bell, B. K. Hamre, and R. C. Pianta. 2013. "Effect of Observation Mode on Measures of Secondary Mathematics Teaching." *Educational and Psychological Measurement* 73(5): 757–783. doi: 10.1177/0013164413486987

Cash, A. H., B. K. Hamre, R. C. Pianta, and S. S. Myers. 2012. "Rater Calibration When Observational Assessment Occurs at Large Scale: Degree of Calibration

and Characteristics of Raters Associated with Calibration." *Early Childhood Research Quarterly* 27(3): 529–542. doi: 10.1016/j.ecresq.2011.12.006

Cohen, J. 1968. "Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit." *Psychological Bulletin* 70(4): 213–220. doi: 10.1037/h0026256

Cronbach, L. J., G. C. Gleser, H. Nanda, and N. Rajaratnam. 1972. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles.* New York, NY: Wiley.

Curby, T. W., P. Johnson, A. J. Mashburn, and L. Carlis. 2016. "Live Versus Video Observations: Comparing the Reliability and Validity of Two Methods of Assessing Classroom Quality." *Journal of Psychoeducational Assessment* 34(8): 765–781. doi: 10.1177/0734282915627115

Curby, T. W., S. E. Rimm-Kaufman, and C. C. Ponitz. 2009. "Teacher–Child Interactions and Children's Achievement Trajectories across Kindergarten and First Grade." *Journal of Educational Psychology* 101(4): 912−925. doi: 10.1037/a0016647

Dobbs-Oates, J., J. N. Kaderavek, Y. Guo, and L. M. Justice. 2011. "Effective Behavior Management in Preschool Classrooms and Children's Task Orientation: Enhancing Emergent Literacy and Language Development." *Early Childhood Research Quarterly* 26(4): 420–429. doi: 10.1016/j.ecresq.2011.02.003

Downer, J. T., M. L. Lopez, K. Grimm, A. Hamagami, R. C. Pianta, and C. Howes. 2012. "Observations of Teacher-Child Interactions in Classrooms Serving Latinos and Dual Language Learners: Applicability of the Classroom Assessment Scoring System in Diverse Settings." *Early Childhood Research Quarterly* 27(1): 21–32. doi: 10.1016/j.ecresq.2011.07.005

Gitomer, D. H., C. A. Bell, Y. Qi, D. F. McCaffrey, B. K. Hamre, and R. C. Pianta. 2014. "The Instructional Challenge in Improving Teaching Quality: Lessons from a Classroom Observation Protocol." *Teachers College Record* 116: 1–32. doi: 10.1177/016146811411600607

Hamre, B. K., B. Hatfield, R. Pianta, and F. Jamil. 2014. "Evidence for General and Domain-Specific Elements of Teacher–Child Interactions: Associations with Preschool Children's Development." *Child Development* 85(3): 1257–1274. doi: 10.1111/cdev.12184

Hamre, B. K., R. C. Pianta, J. T. Downer, J. DeCoster, A. J. Mashburn, S. Jones, et al. 2013. “Teaching through Interactions—Testing a Developmental Framework for Understanding Teacher Effectiveness in Over 4,000 U.S. Early Childhood and Elementary Classrooms.” *The Elementary School Journal* 113(4): 461–487. doi: 10.1086/669616

Hill, H. C., C. Y. Charalambous, and M. A. Kraft. 2012. “When Rater Reliability Is Not Enough: Teacher Observation Systems and a Case for the Generalizability Study.” *Educational Researcher* 41(2): 56–64. doi: 10.3102/0013189X12437203

Hu, B. Y., X. Fan, C. Gu, and N. Yang. 2016. “Applicability of the Classroom Assessment Scoring System in Chinese Preschools Based on Psychometric Evidence.” *Early Education and Development* 27(5): 1–21. doi: 10.1080/10409289.2016.1113069

Jensen, B., M. G. Pérez Martínez, A. M. G. Medina, J. F. Martínez, C. B. Cox, and R. Larsen. 2020. “An Ecological Analysis of the Classroom Assessment Scoring System in K-1 Mexican Classrooms.” *Early Years* 40(4-5): 514–533. doi: 10.1080/09575146.2020.1749035

Leyva, D., C. Weiland, M. Barata, H. Yoshikawa, C. Snow, and E. Treviño. 2015. “Teacher–Child Interactions in Chile and Their Associations with Prekindergarten Outcomes.” *Child Development* 86(3): 781–799. doi: 10.1111/cdev.12342

Li, H., J. Liu, and C. V. Hunter. 2020. “A Meta-Analysis of the Factor Structure of the Classroom Assessment Scoring System (CLASS).” *Journal of Experimental Education* 88(2): 265–287. doi: 10.1080/00220973.2018.1551184

Malmberg, L-E., H. Hagger, K. Burn, T. Mutton, and H. Colls. 2010. “Observed Classroom Quality during Teacher Education and Two Years of Professional Practice.” *Journal of Educational Psychology* 102(4): 916–932. doi: 10.1037/a0020920

Mantzicopoulos, P., B. F. French, H. Patrick, J. S. Watson, and I. Ahn. 2018. “The Stability of Kindergarten Teachers’ Effectiveness: A Generalizability Study Comparing the Framework for Teaching and the Classroom Assessment Scoring System.” *Educational Assessment* 23(1): 24–46. doi: 10.1080/10627197.2017.1408407

Mashburn, A. J., J. T. Downer, S. Rivers, M. Brackett, and A. Martinez. 2014. "Improving the Power of an Experimental Study of a Social and Emotional Learning Program: Application of Generalizability Theory to the Measurement of Classroom-Level Outcomes." *Prevention Science* 15(2): 146–155. doi: 10.1007/s11121-012-0357-3

Mashburn, A., J. P. Meyer, J. Allen, and R. Pianta. 2014. "The Effect of Observation Length and Presentation Order on the Reliability and Validity of Classroom Observations." *Educational and Psychological Measurement* 74(3): 400–422. doi: 10.1177/0013164413515882

Mashburn, A. J., R. B. Pianta, B. K. Hamre, J. T. Downer, O. A. Barbarin, M. Burchinal, et al. 2008. "Measures of Classroom Quality in Prekindergarten and Children's Development of Academic, Language, and Social Skills." *Child Development* 79(3): 732–749. doi: 10.1111/j.1467-8624.2008.01154.x

McCaffrey, D. F., K. Yuan, T. D. Savitsky, J. R. Lockwood, and M. O. Edelen. 2015. "Uncovering Multivariate Structure in Classroom Observations in the Presence of Rater Errors." *Educational Measurement: Issues and Practice* 34(2): 34–46. doi: 10.1111/emip.12061

McHugh, M. L. 2012. "Interrater Reliability: The Kappa Statistic." *Biochem Med (Zagreb)* 22 (3): 276–282.

Merritt, E. G., S. B. Wanless, S. E. Rimm-Kaufman, C. Cameron, and J. L. Peugh. 2012. "The Contribution of Emotional Support to Children's Social Behaviors and Self-Regulatory Skills in First Grade." *School Psychology Review* 41(2): 141–159. doi: 10.1080/02796015.2012.12087517

Pianta, R. C., and B. K. Hamre. 2009. "Conceptualization, Measurement, and Improvement of Classroom Processes: Standardized Observation Can Leverage Capacity." *Educational Researcher* 38(2): 109–119. doi: 10.3102/0013189X09332374

Pianta, R. C., B. K. Hamre, and S. L. Mintz. 2012. *The CLASS-Secondary manual*. Charlottesville: University of Virginia.

Pianta, R. C., K. M. La Paro, and B. K. Hamre. 2008. *The Classroom Assessment Scoring System: Manual*, *Pre-K*. Baltimore, MD: Paul H. Brookes.

Ponitz, C., S. E. Rimm-Kaufman, L. L. Brock, and L. Nathanson. 2009. "Early Adjustment, Gender Differences, and Classroom Organizational Climate in First Grade." *Elementary School Journal* 110(2): 142–162. doi: /10.1086/605470

Sandilos, L. E., and J. C. DiPerna. 2014. “Measuring Quality in Kindergarten Classrooms: Structural Analysis of the Classroom Assessment Scoring System (CLASS K–3).” *Early Education and Development* 25(6): 894–914. doi: 10.1080/10409289.2014.883588

Stuck, A., G. Kammermeyer, and S. Roux. 2016. “The Reliability and Structure of the Classroom Assessment Scoring System in German Preschools.” *European Early Childhood Education Research Journal* 24(6): 873–894. doi: 10.1080/1350293X.2016.1239324

Styck, K. M., C. J. Anthony, L. E. Sandilos, and J. C. DiPerna. 2021. “Examining Rater Effects on the Classroom Assessment Scoring System.” *Child Development* 92(3): 976–993. doi: 10.1111/cdev.13460

von Suchodoletz, A., A. Fäsche, C. Guntzenhauser, and B. K. Hamre. 2014. “A Typical Morning in Preschool: Observations of Teacher–Child Interactions in German Preschools.” *Early Childhood Research Quarterly* 29(4): 509–519. doi: 10.1016/j.ecresq.2014.05.010

Yates, G. C. R., and S. M. Yates. 1990. “Teacher Effectiveness Research: Towards Describing User-Friendly Classroom Instruction.” *Educational Psychology* 10(3): 225–238.

Table 1. Domains, Dimensions and Indicators of the CLASS Pre-K Instrument

Domain	Dimension	Indicators	
		Used Version	Current Version
Emotional Support			
	Positive Climate		
		Relationships	Relationships
		Positive Affect	Positive Affect
		Respect	Respect
		Peer Interaction*	Positive Communication*
	Negative Climate		
		Negative Affect	Negative Affect
		Punitive Control	Punitive Control
		Sarcasm/Disrespect	Sarcasm/Disrespect
		Negativity Not Contained to Events*	-
		Negativity Escalates*	-
		Severe Negativity	Severe Negativity
	Teacher Sensitivity		
		Responsiveness	Responsiveness

	Notices Student's Need for Assistance	Awareness
	Appropriate Activities*	-
	Addresses Problems	Addresses Problems
	Students Seek Support*	-
	Student Comfort	Student Comfort
Regard for Student Perspectives		
	Flexibility and Student Focus	Flexibility and Student Focus
	Support of Autonomy	Support of Autonomy and Leadership
	Student Expression	Student Expression
	Student Responsibility*	-
	Peer Interaction Encouraged*	-
	Restriction of Movement	Restriction of Movement
Classroom Organization		
Behavior Management		
	Proactive	Proactive
	Monitoring*	-
	Redirecting Misbehavior	Redirection of Misbehavior
	Clear Behavioral Expectations	Clear Behavior Expectations

	Loss of Time*	-
	Effective Praise*	-
	Student Misbehavior	Student Behavior
Productivity		
	Provision of Activities*	Maximizing Learning Time*
	Routines	Routines
	Transitions	Transitions
	Preparation	Preparation
	Disruptions*	-
	Managerial Tasks*	-
Instructional Learning Formats		
	Utilization of Materials*	Clarity of Learning Objectives*
	Student Engagement	Student Interest
	Teacher Facilitation	Effective Facilitation
	Modalities	Variety of Modalities and Materials
Instructional Support		
Concept Development		
Higher Order Thinking & Cognition vs. Rote Learning*		-

Analysis and Reasoning	Analysis and Reasoning
Hypothesis Testing*	Creating*
Integration with Previous Concept	Integration
Connections to Real World	Connections to the Real World
Quality of Feedback	
Process Feedback*	Scaffolding*
Feedback Loops	Feedback Loops
Specific Feedback*	Prompting Thought Processes*
Providing Hints	Providing Hints
-	Encouragement and Affirmation*
Language Modeling	
Frequent Conversations	Frequent Conversations
Initiated Language*	-
Open-Ended Questions	Open-Ended Questions
Repetition and Extension	Repetition and Extension
Self and Parallel Talk	Self- and Parallel Talk
Advanced Language	Advanced Language
Student Engagement	-

Active vs. Passive Engagement*	-
Sustained Engagement*	-

Note * difference between the used and the current version of CLASS Pre-K

Table 2. Teacher Demographics

	n	*M* /Percent
Master in Education	62	37.10%
Elementary School Premises	65	44.62%
Female	65	95.38%
Experience More than 15 Years	61	45.90%
Number of Students	64	14.31

Table 3. Indicators of Emotional Support

Positive Climate	n	Teacher (T)	Observer (O)	Episode (E)	T × E	O × E	T × O	Average exact agreement	Min	Max	Weighted Kappa
Relationships	807	0.66	0.03	0.00	0.29	0.01	0.00	0.91	0.73	1.00	0.524
Positive affect	805	0.74	0.03	0.00	0.22	0.00	0.00	0.84	0.50	1.00	0.491
Respect	803	0.64	0.04	0.01	0.28	0.00	0.04	0.89	0.70	1.00	0.502
Peer Interaction	797	0.50	0.07	0.00	0.33	**0.04**	0.07	0.84	0.50	1.00	0.422
Average $\bar{x}$		0.64	0.04	0.00	0.28	0.01	0.03				
Negative Climate	n										
Negative Affect	779	0.34	0.06	0.00	0.52	0.00	0.08	0.95	0.67	1.00	0.489
Punitive Control	778	0.43	0.05	0.01	0.37	**0.07**	0.07	0.95	0.63	1.00	0.360
Sarcasm/Disrespect	778	0.39	0.12	0.00	0.38	0.00	0.11	0.95	0.80	1.00	0.284
Negativity Not Contained to Events	776	0.20	0.14	0.00	0.46	**0.07**	0.14	0.95	0.78	1.00	0.268
Escalating Negativity	776	0.21	0.04	0.00	0.50	0.02	**0.23**	0.96	0.76	1.00	0.297
Severe Negativity	771	0.00	0.12	0.00	0.47	0.00	**0.40**	0.98	0.88	1.00	0.081
Average $\bar{x}$		0.26	0.09	0.00	0.45	0.03	0.17				

Teacher Sensitivity	n										
Responsive	833	0.63	0.12	0.01	0.22	0.01	0.02	0.86	0.50	1.00	0.368
Notices Student's Need for Assistance	819	0.64	0.06	0.01	0.22	0.00	0.07	0.86	0.56	1.00	0.410
Appropriate Activities	823	0.63	0.02	0.00	0.29	0.00	0.06	0.90	0.56	1.00	0.269
Addresses Problems	796	0.64	0.10	0.00	0.21	0.01	0.05	0.85	0.50	1.00	0.336
Students Seek Support	816	0.73	0.01	0.00	0.19	0.00	0.07	0.82	0.33	1.00	0.373
Student Comfort	812	0.66	0.06	0.00	0.18	0.01	0.08	0.88	0.63	1.00	0.327
Average $\overline{x}$		0.66	0.06	0.00	0.22	0.01	0.06				
Regard for Student Perspectives	n										
Flexibility & Student Focus	835	0.52	0.03	0.00	0.40	0.00	0.06	0.77	0.00	1.00	0.450
Support of Autonomy	831	0.38	0.10	0.00	0.48	0.00	0.05	0.75	0.50	1.00	0.394
Student Expression	834	0.41	0.06	0.00	0.52	0.01	0.00	0.77	0.00	1.00	0.468
Student Responsibility	827	0.41	0.02	0.00	0.41	0.03	0.12	0.73	0.30	1.00	0.359
Peer Interaction Encouraged	825	0.37	0.04	0.00	0.57	0.00	0.03	0.75	0.00	1.00	0.412

Restriction of Movement	830	0.27	0.12	0.01	0.55	0.00	0.06	0.70	0.00	1.00	0.408
Average $\bar{x}$		0.39	0.06	0.00	0.49	0.01	0.05				

Note: Average exact agreement = the proportion of times the observers have rated each item within 1 point from each other. Min = the lowest exact agreement (proportion) across observer-pairs. Max = the largest exact agreement (proportion) across observer-pairs. Values in bold are outside the third quartile plus 1.5 × the interquartile range for the observer, O × T and O × E interactions.

Table 4. Indicators of Classroom Organization

Behavior Management	n	Teacher (T)	Observer (O)	Episode (E)	T × E	O × E	T × O	Average exact agreement	Min	Max	Weighted Kappa
Proactive	830	0.51	0.02	0.00	0.31	**0.04**	0.11	0.83	0.56	1.00	0.398
Monitoring	830	0.54	0.06	0.01	0.30	0.00	0.10	0.83	0.44	1.00	0.357
Redirecting Misbehavior	818	0.66	0.00	0.00	0.25	0.00	0.09	0.82	0.00	1.00	0.421
Clear Behavioral Expectations	826	0.52	0.04	0.01	0.27	0.00	**0.16**	0.85	0.50	1.00	0.331
Loss of Time	822	0.52	0.02	0.00	0.36	**0.03**	0.07	0.75	0.00	1.00	0.366
Effective Praise	796	0.57	0.06	0.00	0.29	0.00	0.07	0.79	0.44	1.00	0.381
Student Misbehavior	828	0.45	0.00	0.00	0.47	0.00	0.07	0.80	0.44	1.00	0.432
Average $\bar{x}$		0.54	0.03	0.00	0.32	0.01	0.10				
Productivity	n										
Provision of Activities	832	0.21	0.02	0.00	0.63	0.02	0.11	0.91	0.72	1.00	0.305
Routines	831	0.34	0.05	0.01	0.50	0.00	0.09	0.92	0.80	1.00	0.360
Transitions	814	0.35	0.10	0.03	0.40	0.00	0.12	0.86	0.40	1.00	0.358
Preparation	815	0.42	0.06	0.00	0.47	0.01	0.05	0.91	0.78	1.00	0.351

Disruptions	822	0.49	0.00	0.00	0.43	0.00	0.08	0.84	0.00	1.00	0.402
Managerial Tasks	800	0.18	**0.30**	0.00	0.41	0.02	0.10	0.83	0.50	1.00	0.320
Average $\overline{x}$		0.33	0.09	0.01	0.47	0.01	0.09				
Instructional Learning Formats	n										
Utilization of Materials	801	0.29	0.06	0.03	0.60	0.00	0.02	0.84	0.38	1.00	0.439
Student Engagement	804	0.33	0.07	0.00	0.57	0.00	0.03	0.83	0.60	1.00	0.293
Teacher Facilitation	797	0.33	0.07	0.08	0.52	0.00	0.00	0.82	0.50	1.00	0.367
Modalities	788	0.35	0.13	0.03	0.46	0.00	0.03	0.74	0.39	1.00	0.355
Average $\overline{x}$		0.32	0.08	0.03	0.54	0.00	0.02				

Note. Average exact agreement = the proportion of times the observers have rated each item within 1 point from each other. Min = the lowest exact agreement (proportion) across observer-pairs. Max = the largest exact agreement (proportion) across observer-pairs. Values in bold are outside the third quartile plus 1.5 × the interquartile range for the observer, O × T and O × E interactions.

Table 5. Indicators of Instructional Support and Student Engagement

Concept Development	n	Teacher (T)	Observer (O)	Episode (E)	T × E	O × E	T × O	Average exact agreement	Min	Max	Weighted Kappa
Higher Order Thinking & Cognition vs. Rote Learning	657	0.29	0.11	0.04	0.48	0.00	0.08	0.79	0.50	1.00	0.417
Analysis and Reasoning	654	0.28	0.10	0.06	0.54	0.01	0.02	0.80	0.50	1.00	0.423
Hypothesis Testing	643	0.24	0.17	0.05	0.48	0.00	0.07	0.69	0.00	1.00	0.364
Integration with Previous Concept	644	0.24	0.15	0.03	0.45	0.00	0.13	0.89	0.60	1.00	0.352
Connections to Real World	650	0.24	0.11	0.06	0.51	0.00	0.08	0.77	0.47	1.00	0.399
Average $\bar{x}$		0.26	0.13	0.04	0.49	0.00	0.08				
Quality of Feedback	n										
Process Feedback	817	0.46	0.16	0.04	0.31	0.00	0.03	0.71	0.00	1.00	0.361
Feedback Loops	813	0.57	**0.32**	0.03	0.00	0.00	0.08	0.73	0.44	1.00	0.340
Specific Feedback	816	0.39	0.13	0.05	0.34	**0.03**	0.06	0.70	0.00	1.00	0.330

Providing Hints	780	0.43	0.07	0.06	0.36	0.00	0.07	0.74	0.52	1.00	0.359
Average $\bar{x}$		0.46	0.17	0.04	0.25	0.01	0.06				
Language Modeling	n										
Frequent Conversations	832	0.40	0.08	0.04	0.42	0.00	0.06	0.77	0.33	1.00	0.416
Initiated Language	829	0.55	0.08	0.01	0.30	0.00	0.06	0.73	0.50	1.00	0.388
Open-Ended Questions	821	0.41	0.09	0.01	0.42	0.01	0.06	0.79	0.50	1.00	0.409
Repetition and Extension	811	0.35	0.14	0.05	0.37	0.01	0.08	0.67	0.33	1.00	0.350
Self and Parallel Talk	812	0.35	**0.28**	0.03	0.28	0.00	0.06	0.61	0.25	1.00	0.315
Advanced Language	824	0.27	0.23	0.06	0.34	0.00	0.10	0.73	0.33	1.00	0.348
Average $\bar{x}$		0.39	0.15	0.03	0.35	0.01	0.07				
Student Engagement	n										
Active vs. Passive Engagement	785	0.37	0.00	0.03	0.51	0.00	0.09	0.74	0.00	1.00	0.410
Sustained Engagement	787	0.31	0.09	0.01	0.48	0.00	0.10	0.65	0.00	0.93	0.330
Average $\bar{x}$		0.34	0.05	0.02	0.50	0.00	0.09				

Note. Average exact agreement = the proportion of times the observers have rated each item within 1 point from each other. Min = the lowest exact agreement (proportion) across observer-pairs. Max = the largest exact agreement (proportion) across observer-pairs. Values in bold are outside the third quartile plus 1.5 × the interquartile range for the observer, T × O and O × E interactions.