Journal of Applied Developmental Psychology 48 (2017) 79–91

Contents lists available at ScienceDirect

Journal of Applied Developmental Psychology

Measurement properties of the CLASS Toddler in ECEC in The Netherlands

Pauline L. Slot *, Jan Boom, Josje Verhagen, Paul P.M. Leseman

Utrecht University, the Netherlands

ARTICLE INFO

Article history:
Received 10 October 2014
Received in revised form 17 November 2016
Accepted 30 November 2016
Available online 12 January 2017

Keywords:
Early childhood education and care
Classroom quality
Classroom Assessment Scoring System Toddler
Item properties
Construct and criterion validity

ABSTRACT

The present study investigated the measurement properties of the Classroom Assessment Scoring System (CLASS) Toddler using data from 276 classrooms and 375 teachers in Dutch early childhood education and care provisions. First, confirmatory factor analyses based on the CLASS Toddler indicators confirmed the eight-dimension structure of the instrument. Second, a three-domain structure, involving Emotional Support, Behavioral Support and Engaged Support for Learning, fitted the data better than the commonly found two-domain structure that combines Emotional and Behavioral Support into one domain. Third, using Item Response Theory (IRT), we found good item difficulty and discrimination parameters of the CLASS Toddler indicators. Finally, significant associations were found between the CLASS Toddler domains and children-to-staff ratio, teachers' work experience and provided play and literacy activities, supporting the criterion validity of the instrument. Taken together, these results showed adequate measurement quality of the CLASS Toddler. Implications for practice are discussed.

1. Introduction

Observation instruments assessing the quality of care and education provided in early childhood settings are extensively used in evaluation research, as well as in professionalization programs at the center level and in quality monitoring and quality improvement at the system level (Martinez-Beck, 2011). Examples of the latter are the Quality Rating and Improvement Systems (QRIS) in the United States (US; Tout et al., 2010) and the Caregiver Interaction Profiles (CIP; Helmerhorst, Riksen-Walraven, Vermeer, Fukkink, & Tavecchio, 2014) in The Netherlands (Brancheorganisatie Kinderopvang, 2014). Given the increased use of quality measurement instruments in high stakes contexts, with often profound consequences for accountability and licensing of service providers, adequate measurement quality of these instruments is essential (Hestenes et al., 2015). A key question is if these instruments measure ECEC quality in a reliable and valid way and tap the construct of quality as theoretically intended.

Several scholars have recently suggested that widely used observation instruments assessing ECEC quality, such as the Early Childhood Environmental Rating Scale Revised (ECERS-R) and the Classroom Assessment Scoring System (CLASS), may not tap quality adequately (Burchinal, Kainz, & Cai, 2011; Colwell, Gordon, Fujimoto, Kaestner, & Korenman, 2013; Gordon, Fujimoto, Kaestner, Korenman, & Abner, 2013; Layzer & Goodson, 2006; Zaslow et al., 2010). This might be due to flaws in the measurement quality of these instruments. In fact, the few studies to date that have addressed the measurement properties of these instruments have revealed several problems regarding the scaling of the items, the structural validity, evaluated by examining the factor structure, and the criterion validity (Cassidy, Hestenes, Hegde, Hestenes, & Mims, 2005; Colwell et al., 2013; Cryer, Tietze, Burchinal, Leal, & Palacios, 1999; De Kruif, McWilliam, Maher Ridley, & Wakely, 2000; Gordon et al., 2013; Perlman, Zellman, & Le, 2004). This poses a threat to the reliability and validity of the results obtained with these instruments and, consequently, to the fairness of decisions based on these instruments (Cassidy et al., 2005; Colwell et al., 2013; Perlman et al., 2004). Moreover, research on the measurement properties of quality assessment instruments in ECEC has been mostly conducted within a classical test theory (CTT) framework, using correlational and factor analyses (Gordon et al., 2013). The CTT framework, however, has a number of limitations. One limitation is that the CTT approach only allows for an evaluation of a measure as a whole, disregarding more detailed information at the indicator or item level. Another is that, in CTT, the measurement properties of the instrument and the ability of the sample under study cannot be separated (de Ayala, 2013; Hambleton & Jones, 1993). Item response theory (IRT) provides alternatives to both issues as will outlined in more detail below.

To summarize, there is a clear need for a thorough examination of the measurement properties of instruments for assessing the quality of ECEC. In the current study, we investigate a recently developed and already widely used ECEC quality observation instrument, the CLASS Toddler.

* Corresponding author at: Department of Child, Family, and Education Studies, Utrecht University, the Netherlands.
E-mail address: P.L.Slot@uu.nl (P.L. Slot).

http://dx.doi.org/10.1016/j.appdev.2016.11.008

More specifically, we examine the measurement properties of this instrument in Dutch ECEC settings by looking at three aspects of validity: (i) structural validity (i.e., factor structure), (ii), item characteristics of the behavioral indicators (i.e., indicator difficulty and discrimination parameters), and finally, (iii), criterion validity. In investigating these aspects, the current study combines a CTT with an IRT approach. These approaches together can provide more detailed and comprehensive information on the measurement quality of the CLASS Toddler.

1.1. Measuring teacher-child interactions

The CLASS is theoretically grounded in well-supported developmental and educational theories, and focuses on proximal classroom processes (Hamre & Pianta, 2007). The CLASS framework is based on the assumption that core characteristics of effective teacher-child interactions show continuity across age levels and across classrooms, from infant/toddler classrooms through high school, although their exact behavioral manifestations may differ by children's age (Hamre et al., 2013). The CLASS Pre-K was the first version of the CLASS and provided the basis for the development of other versions. Hence, in this section, we start by discussing the overarching theoretical framework of the CLASS and empirical evidence on the measurement properties of the CLASS Pre-K, before turning to the CLASS Toddler.

Classroom *process quality* is defined as the quality of the social, emotional, physical, and instructional aspects of children's day-to-day experiences in classrooms, with a particular focus on children's interactions with teachers, peers, and materials (Howes et al., 2008). Classroom processes are viewed as the primary drivers of children's development and learning in educational settings (cf. Bronfenbrenner & Morris, 2006), as outlined in the Teaching through Interactions framework (for an overview, see Hamre et al., 2013). This framework was further developed in observational studies in over 4000 early childhood and elementary school classrooms, evaluating the quality of proximal classroom interactions along ten dimensions, divided in three broad domains: *Emotional Support*, *Classroom Organization*, and *Instructional Support*.

The domain Emotional Support consists of the dimensions *Positive Climate*, *Negative Climate*, *Teacher Sensitivity*, and *Regard for Child Perspectives*. Two theories have provided the basis for evaluating emotional support in classrooms within the overarching CLASS framework, specifically attachment theory (Ainsworth, Blehar, Waters, & Wall, 1978; Bowlby, 1969) and self-determination theory (Ryan & Deci, 2000). These theories emphasize the importance of positive and affectionate relationships between children and adults that provide children with a secure base from which they can explore their environment and develop a sense of autonomy and competence (Sroufe, 2000). The domain Classroom Organization consists of the dimensions *Behavior Management*, *Productivity*, and *Instructional Learning Formats*. The evaluation of classroom processes within this domain specifically focuses on teachers' strategies in developing classroom routines, organizing developmentally appropriate activities and managing children's engagement in these activities, and is informed by research on children's self-regulation (Blair, 2003; Raver, 2004) and teachers' effective classroom management (Emmer & Strough, 2001). Finally, the domain Instructional Support consists of the dimensions *Concept Development*, *Quality of Feedback*, and *Language Modeling*. The evaluation of instructional interactions is based on research into children's cognitive and language development (e.g., Catts, Fey, Zhang, & Tomblin, 1999; Taylor, Pearson, Peterson, & Rodriguez, 2003), and focuses on teachers' strategies in promoting meaningful learning opportunities by scaffolding children's learning during activities, providing performance-related feedback, encouraging higher-order thinking, and modeling advanced language use (Davis & Miyake, 2004; Mayer, 2002; Skibbe, Behnke, & Justice, 2004).

The newer CLASS Toddler (La Paro, Hamre, & Pianta, 2011) is also based on this theoretical framework. Positive and affectionate relations with adults are deemed essential for establishing secure attachments with teachers, providing the basis for young children to develop a sense of autonomy while maintaining relationships with significant adults, such as teachers, and consequently allowing for exploration of their environment and engaging in peer interactions (Bronson, 2000; Sroufe, 2000). Toddlerhood is a developmental period in which children gradually learn how to behave and to adjust according to classroom expectations, how to regulate their emotions and how to interact with peers in play activities (Bronson, 2000; Kopp, 1982). Explicit learning goals requiring specific learning formats, classroom management and the productive use of time, as in (pre)kindergarten classrooms, is usually not yet at stake. Therefore, the CLASS Toddler lacks a separate domain referring to these aspects of quality and conceptualizes teacher-child interactions along eight dimensions captured in two instead of three broad domains (see Fig. 1).

The domain *Emotional and Behavioral Support* of the CLASS Toddler reflects teachers' responsiveness to children's needs by creating warm and affectionate relationships and by supporting children's autonomy and self-regulation. This domain consists of the dimensions *Positive Climate*, *Negative Climate*, *Teacher Sensitivity*, *Regard for Child Perspectives*, and *Behavior Guidance*. The domain *Engaged Support for Learning* encompasses the teachers' ability to engage in activities and interactions with children that facilitate their learning and development. This domain consists of the dimensions *Facilitation of Learning and Development*, *Quality of Feedback*, and *Language Modeling*. The dimension Facilitation of Learning and Development combines both children's active engagement, in the Pre-K version captured under Instructional Learning Formats, and the teacher's role in supporting cognitive development, in the Pre-K version captured under Concept Development. There is emerging evidence showing that classroom quality as assessed with the CLASS Toddler is positively related to children's social-emotional and language development (Bandel, Aikens, Vogel, Boller, & Murphy, 2014; La Paro, Williamson, & Hatfield, 2014).

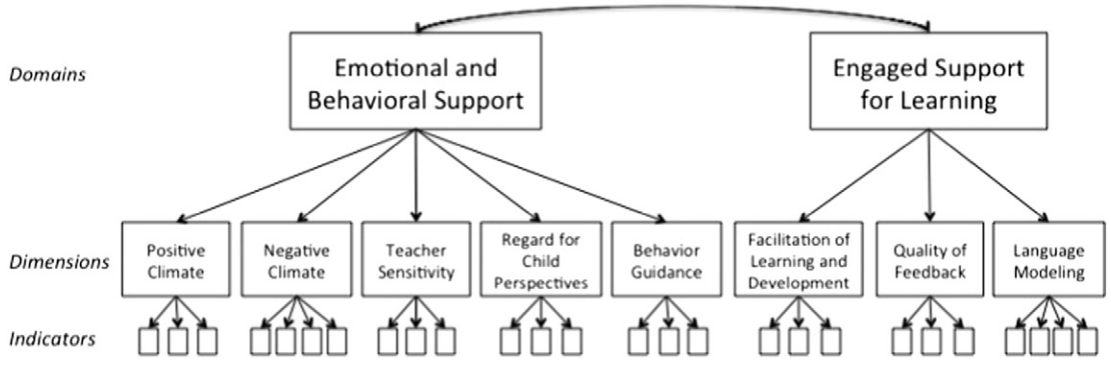

Fig. 1. Conceptual model of the CLASS Toddler, based on the CLASS Toddler manual.

1.2. Structural validity of the CLASS Toddler: domain and dimensional structure

1.2.1. Domain structure

The structural validity of the CLASS Pre-K has been extensively investigated in studies in the US as well as in other countries. Although the findings from studies outside the US generally support findings in the US, there are indications of cultural differences, as will be reviewed below. For the newer CLASS Toddler, the evidence on the measurement quality is still limited and pertains to the US only. In addition, previous research on the CLASS Pre-K and Toddler has solely focused on evaluating the domain structure, while disregarding the underlying structure of behavioral indicators and dimensions, which, as will be argued in the following section, constitutes a gap in our current knowledge on the measurement properties of the CLASS.

Concerning the CLASS Pre-K, Hamre et al. (2013), in their study on preschool and elementary school classrooms until the fifth grade, showed that a three-domain structure as proposed in the Teaching through Interaction model was preferred over a two-domain structure (referred to as the Social and Instructional model) or a one-domain structure (referred to as the Effective Teaching model). This three-domain structure has been replicated in research on preschools in countries other than the US as well (Cadima, Leal, & Burchinal, 2010; Leyva et al., 2015; Pakarinen et al., 2010; von Suchodoletz, Fäsche, Gunzenhauser, & Hamre, 2014). However, one of the dimensions of the Emotional Support domain, Negative Climate, appeared to function differently in other countries than the US (e.g., Leyva et al., 2015; Pakarinen et al., 2010; von Suchodoletz et al., 2014). For example, a Finnish study on the CLASS Pre-K found that the Negative Climate dimension, originally part of the Emotional Support domain, did not fit well into the three-domain model. Rather, this dimension loaded on Classroom Organization, which led the researchers to exclude the dimension from the model altogether (Pakarinen et al., 2010). Furthermore, in studies on the CLASS Pre-K in Chile and Germany, the Negative Climate dimension correlated stronger with the Behavior Guidance dimension, which is part of the Classroom Organization domain, than with the other dimensions of the Emotional Support domain (Leyva et al., 2015; von Suchodoletz et al., 2014), suggesting that cultural differences may be at stake.

To date, the structural validity of the CLASS Toddler has been investigated in only three studies, which were all conducted in the US (Bandel et al., 2014; Castle et al., 2016; La Paro et al., 2014). These studies confirmed the proposed two-domain structure of the instrument. The studies involved 93 licensed center-based childcare programs in North Carolina (La Paro et al., 2014) and over 200 Early Head Start programs across the US (Bandel et al., 2014; Castle et al., 2016). Bandel et al. (2014) also investigated a three-domain solution, because they hypothesized that the dimension Behavior Guidance might load onto a separate factor, as with the CLASS Pre-K, but their factor analysis did not support this. Whether the two-domain structure of the CLASS Toddler is also valid in other countries than the US, still needs to be established, however. For example, the pattern of correlations found for Negative Climate and Behavior Guidance in Finland, Germany and Chile suggests that these dimensions may constitute a separate domain: Behavioral Support (Leyva et al., 2015; Pakarinen et al., 2010; von Suchodoletz et al., 2014). Also from a theoretical point of view this seems plausible as both dimensions capture aspects of children's behavior and the way teachers respond to this behavior. In a classroom led by teachers with clear behavioral expectations and positive reinforcement of desirable behavior, children may show less misbehavior and negativity, and teachers, in turn, may show less negativity in their interactions with children. Since a third domain consisting of Negative Climate and Behavior Guidance seems plausible from both an empirical and a theoretical point of view, we employed confirmatory factor analyses to compare the originally proposed two-domain structure with a three-domain structure (Emotional Support, Behavioral Support and Engaged Support for Learning model). Following the earlier studies reviewed above, we also compared the two-domain structure with a one-domain structure (effective teaching model).

1.2.2. Dimensional structure

The CLASS Toddler has a hierarchical structure (see Fig. 1) representing, at the highest level, the two domains that are based on the eight dimensions, which in turn are based on three or four behavioral indicators per dimension. The indicators specify the concrete and observable behaviors of teachers that observers can rate during the selected observation periods. They help observers to assign a score to the dimensions (La Paro et al., 2011), which are then averaged over dimensions into a domain score. Earlier research on the structural validity of the CLASS Toddler has focused solely on the domain structure; the dimensional structure of the CLASS Toddler indicators was not evaluated. However, whether pooling the scores on the indicators into a dimension score is warranted, still needs to be established.

In order to base the quality assessment on a representative set of classroom situations, the general CLASS observation procedure prescribes to divide the total observation time per classroom into four separate observation cycles. This yields a nested data structure calling for appropriate analysis methods. Simply aggregating the detailed observation measures to the classroom level, as is common in ECEC research, may lead to the loss of potentially relevant information (Hox, 2010). Moreover, it is uncertain whether the same construct is measured at both levels. To date, studies investigating the measurement properties of the CLASS have not investigated whether the factor structure at the cycle level, reflecting variation within classrooms, is similar to the factor structure based on the aggregated measurements at the classroom level, reflecting the variation between classrooms. In the current study, we employed a multilevel approach to investigate the structure of the CLASS Toddler at both the within-classroom level, representing variation within the classroom, and at the between-classroom level, reflecting variation between classrooms.

1.3. Item properties of the CLASS Toddler indicators

Previous studies on the CLASS Toddler have only examined the measurement properties of the instrument as a whole. Consequently, information on the measurement properties of the items, that is, the behavioral indicators, is lacking. However, good measurement quality requires the use of an appropriate set of items. A measure can only reliably differentiate in a wide ability range and show sensitivity to differences between teachers both in the lower and higher regions of the ability dimension if items vary in difficulty level. In addition, each item should show sufficient discrimination value, indicating the amount of information each item contributes to the final dimension score. Sufficiently high discrimination values are needed in order to ensure that all relevant aspects are well represented in the final score.

To the best of our knowledge, only two studies have examined the item properties of classroom quality measures, neither of which have included the CLASS Toddler. First, in a large cohort study in the US, the measurement properties of the items of the Arnett Caregiver Interaction Scale (CIS) were investigated (Colwell et al., 2013). This study showed that several items and the scale as a whole did not differentiate well between caregivers in the moderate and high range of provided quality. Likewise, a study into the validity of the ECERS-R revealed disordering of individual indicators within the quality items that were scored, violating the scaling assumptions (Gordon et al., 2013). Disordering of items, in this case, meant that lower actual quality was more likely to receive a higher score, while higher actual quality was more likely to receive a lower score. Gordon et al. (2013) showed that there was at least one disordered indicator in each quality aspect of the ECERS-R. In the current study, we examined the measurement properties of the CLASS Toddler's primary observation categories, the indicators, using IRT.

1.4. Criterion validity: associations between observed classroom quality and classroom and teacher characteristics

Several studies have confirmed the criterion validity of CLASS Pre-K and CLASS Toddler, through expected associations between CLASS ratings and teacher and classroom characteristics (e.g., Castle et al., 2016; Mashburn et al., 2008; Thomason & La Paro, 2009). Specifically, earlier work has shown weak to moderately strong positive relationships between classroom quality and teachers' pre-service qualifications and work experience (Castle et al., 2016; LoCasale-Crouch et al., 2007; Thomason & La Paro, 2009; Vogel, Caronongan, Thomas et al., 2015). Higher quality in toddler classrooms has also been found to be positively related to classroom characteristics, such as smaller group size, more favorable children-to-staff ratios, and a higher proportion of dual language learners (Thomason & La Paro, 2009; Vogel, Caronongan, Thomas et al., 2015; Vogel, Caronongan, Xue et al., 2015). Finally, observed classroom quality, as observed with the CLASS Pre-K, has been shown to be positively related to the implemented curriculum of literacy and math activities (Cabell et al., 2013; Howes et al., 2008). Evidence for the criterion validity of the CLASS Toddler, however, is still limited and only provided by US studies. Hence, one of the aims of the present study was to investigate the criterion validity of the CLASS Toddler by examining how classroom quality relates to several teacher, classroom and curriculum characteristics of Dutch ECEC.

1.5. Early childhood education and care in The Netherlands

The Dutch ECEC system consists of two main types of provision. The first type is center-based daycare for children from birth to three years of age. This type of daycare is typically attended for, on average, two full days a week by children of working parents (NCKO, 2011). The second type concerns preschools for two- and three-year-old children, which are attended for two to four half days a week. Many preschools and increasingly also daycare centers offer an education program targeted at children from lower socioeconomic backgrounds. The maximum age of the children attending both types of ECEC provisions is three, since, at their fourth birthday, virtually all children in The Netherlands enter the kindergarten department of primary school (Van Tuijl & Leseman, 2007).

The Dutch ECEC system is strongly regulated. The Dutch Childcare Act of 2005 prescribes a children-to-teacher ratio of 7:1 for two- and three-year-old children as well as a maximum group size of 12 for two-year-old children and 16 for three-year-old children (Convenant Kwaliteit Kinderopvang, 2008). Also, teachers are required to have completed a minimum of three -years training in a relevant vocational training program. The OKE (Promoting Development through Quality and Education) Act of 2010 brings daycare centers and preschools under the same statutory quality framework and emphasizes the equal importance of social, emotional and cognitive outcomes for children. Although the two types of ECEC differ in the age range and socioeconomic background of the children served and stem from different traditions in ECEC (with a care and education orientation, respectively), differences in structural and process quality have largely disappeared due to new legislation (Slot, Leseman, Verhagen, & Mulder, 2015).

1.6. The current study

The current study had three aims. The first was to investigate the structural validity, that is, the dimensional and domain structure of the CLASS Toddler, while accounting for the nested data structure of the multiple observation cycles within classrooms. We started with an examination of the proposed dimensional structure of the CLASS Toddler using an IRT approach to account for the ordinal structure of the data provided by the indicators. Next, the overarching domain structure of the CLASS dimensions was investigated, following a CTT approach.

The second aim of our study was to evaluate the measurement properties of the CLASS indicators. Taking the outcomes of the previous analysis step as starting point, we evaluated the indicator difficulty and discrimination values using IRT analyses. Although the quality of a measurement instrument is critically dependent upon the quality of its primary sources of information at the item level, the classical CTT approach provides little information about the scaling of items and relevant item statistics (Bryant, Burchinal, & Zaslow, 2010; Hambleton & Jones, 1993; Lambert, Nelson, Brewer, & Burchinal, 2006). Moreover, in the CTT approach, ability and item parameters are not invariant (de Ayala, 2013; Hambleton & Jones, 1993). Characteristics of the instrument affect a person's ability, that is, the quality a teacher provides, whereas characteristics of the sample, in turn, affect item statistics, such as item difficulty and item discrimination. Therefore, in CTT, the scores derived from a specific measure are only representative of a person's ability on that particular measure, while the item statistics are only valid for the particular sample (de Ayala, 2013; Hambleton & Jones, 1993). IRT proposes an alternative way of evaluating measurement quality based on the invariance at the ability and item parameter level. Regarding quality observation measures, IRT models assume that teachers' observed performance reflects a latent ability, which is independent of the particular items used to assess this ability. In turn, item statistics used in IRT, such as item difficulty and item discrimination, are assumed to be independent of the particular sample in which they are estimated (de Ayala, 2013; Hambleton & Jones, 1993).

Our third aim was to evaluate the criterion validity of the CLASS Toddler by investigating associations between observed process quality, on the one hand, and several structural quality aspects and curriculum, on the other. We expected associations between the CLASS domains and structural quality characteristics, such as group size, children-to-teacher ratio, teacher's education level, work experience, and ethnic-cultural group composition. However, previous research in The Netherlands has shown mostly weak to moderate associations between structural aspects and process quality, presumably due to the strong national regulations concerning structural quality characteristics resulting in restricted variance (de Kruif et al., 2009; Fukkink, Gever Deynoot-Schaub, Helmerhorst, Bollen, & Riksen-Walraven, 2013; Slot et al., 2015; Vermeer et al., 2008). Therefore, we expected the magnitude of the associations to be small to moderate. We also expected weak to moderate positive associations between the CLASS domains and curriculum characteristics, especially between the provision of literacy and math activities and observed process quality (Cabell et al., 2013; Howes et al., 2008). In addition, the provision of child-centered play is considered an important aspect of ECEC quality (Cabell et al., 2013). However, there is also evidence that play is not necessarily related to high quality teacher-child interactions (Singer, Nederend, Penninx, Tajik, & Boom, 2014). Therefore, we took an exploratory approach regarding the relations between play and observed classroom quality.

2. Method

2.1. Participants

The present study used data from the first measurement wave of the national cohort study pre-COOL. Pre-COOL was commissioned by the Dutch Ministry of Education, Culture and Sciences and the National Science Foundation to investigate the effects of preschool education and care provisions in The Netherlands on children's developmental outcomes (www.pre-cool.nl). The study started in 2010, when children were about two years of age. Children were followed up until age 5, with four annual child assessments. At age five, children entered the national cohort study COOL (Cohort Study on Educational Careers; www.cool5-18.nl) on students' careers in primary and secondary education. In COOL they are being followed up until age eighteen. To increase the likelihood of pre-COOL children entering primary schools that take part in COOL, the sample was recruited in the following way. First, a

random sample of 300 primary schools participating in COOL was drawn. Next, the 139 primary schools that agreed to participate (46.3%) were asked to identify the preschools and daycare centers that were attended most by their new students. Municipal records and the internet were used to identify additional preschools and daycare centers in the neighborhoods of the COOL schools. About 500 centers were approached, of which 263 agreed to participate in pre-COOL (52.6%). The participating preschools and daycare centers were geographically well-spread over The Netherlands, located in urban, semi-urban and rural areas, and did not differ significantly on these geographical characteristics from non-participating preschools and daycare centers (Pre-COOL Consortium, 2012). For logistic reasons, observations were only conducted in centers with more than four children who had also participated in the child assessments of pre-COOL. This yielded 162 centers (61.6%) with a total of 276 classrooms, of which 155 classrooms in preschools and 121 classrooms in daycare centers. In addition, for the purpose of investigating criterion validity, we used teacher report data. In total, 375 teachers of 182 centers (69.2% of the total sample) participated in the study by filling out a teacher questionnaire, providing information on 295 classrooms (170 in preschools, 125 in daycare centers). Almost all teachers were women (99.2%) and predominantly Caucasian (89.4%). For the evaluation of the criterion validity of the CLASS Toddler, complete observation data and self-reports were available for 110 classrooms (39.8%). Descriptive information on center and classroom characteristics are shown in Table 1.

2.2. Classroom measures

2.2.1. Observed process quality

The CLASS Toddler (La Paro et al., 2011) was used to assess classroom process quality. At the outset of the study, we piloted both the CLASS pre-K and the CLASS Toddler. The CLASS pre-K did not seem sufficiently informative for the two- to three-year-olds. Particularly the instructional dimensions of the CLASS Pre-K showed severely limited variance. The CLASS Toddler was more sensitive to variation in instructional quality in the toddler classrooms and thus was used in the current study.

Table 1
Descriptive statistics for classroom and teacher characteristics reflecting the aggregated classroom level information.

	N	*M*	*SD*	Range
Children-to-teacher ratio[a]	1083	5.12	2.22	0.33–16
Group size[a]	1083	9.76	3.59	1–25
Professional development activities	294	3.17	0.98	1.22–6.33
Frequency (*N*)/*Percent* (*M*)	*N*	%		
Educational program	238	84.40		
Classroom > 30% non-Dutch speaking children	146	41.40		
Age composition classrooms[b]				
0 years	62	20.80		
1 year	77	25.90		
2 years	293	97.00		
3 years	161	54.00		
Educational level				
Lower preparatory vocational track	35	12.30		
4 years secondary vocational	116	40.80		
1 or 2 years intermediate vocational training	20	7.00		
3 or 4 years intermediate vocational training	50	17.60		
5 years secondary	44	15.50		
6 years secondary[c]	8	2.80		
Higher vocational	5	1.80		
University	6	2.10		
Teacher ethnicity				
Native Dutch	341	88.80		
Immigrant	43	11.20		

[a] Based on the observation cycles.
[b] Percentages reflect the number of classrooms, which included children of a given age as a percentage of the total number of classrooms for which this information was available. As most groups included children of different ages, the sum across percentages adds up to far over 100.
[c] Entry level university.

An officially approved Dutch translation of the CLASS manual was made for the present study (Slot, Leseman, Mulder, & Verhagen, 2013). All observers were trained by a licensed CLASS trainer and achieved at least 80% agreement within one scale point with the CLASS trainer on each dimension on an online test (average agreement was 86.4%; agreement by chance was 33.0%), as recommended by the developers of the CLASS. Following this online test, the trainer conducted live observations with all observers (22 in total), prior to data collection. On average, the observers achieved 89.9% agreement within one scale point on each dimension with the CLASS trainer during the live observations, again following the recommended procedure (all observers achieved at least 80% agreement; agreement by chance was 33.0%). Each classroom was observed during one morning and all classrooms were observed within a three-month period after training. Following the CLASS manual, observers rated classroom processes and teacher behavior during four 15 to 20 min observation cycles on the observation morning.

Classroom quality was rated on eight dimensions, using 7-point scales ranging from 1 or 2 (*classroom is low on that aspect*); to 3, 4 or 5 (*classroom is in the midrange*); and to 6 or 7 (*classroom is high on that aspect*). Following the CLASS manual, two overarching domains were distinguished (La Paro et al., 2011). For the first domain, Emotional and Behavioral Support, the observed processes were evaluated on five dimensions: Positive Climate reflects the warmth, respect, and enjoyment displayed during interactions of the teacher and the children; Negative Climate reflects the overall negativity expressed in the classroom by the teacher and the children (scores are reversed); Teacher Sensitivity stands for the extent to which the teacher is aware of and responsive to children's needs; Regard for Child Perspective captures the degree to which the teacher's interactions with children and classroom activities follow children's interests, and the degree to which children's independence is encouraged. Behavior Guidance refers to the teacher's ability to promote positive behavior and redirect problem behavior. In the domain Engaged Support for Learning, observed processes were evaluated on three dimensions: Facilitation of Learning and Development considers how well the teacher facilitates activities to support children's learning and development; Quality of Feedback assesses the degree to which the teacher's feedback promotes learning and expands children's participation; Language Modeling refers to the extent to which the teacher fosters, models and encourages children's use of language. Descriptive statistics on the CLASS dimensions are presented in Table 2.

The CLASS dimension scores are based on the evaluation of three to four indicators. Each indicator is scored on an ordinal scale with five levels, varying from 'low' (1), 'low/mid' (2), 'mid' (3), 'mid/high' (4), to 'high' (5). A low score is given when the specified behavior is not or hardly observed; a midscore is given when the behavior is sometimes observed; a high score indicates a prevalent occurrence of the behavior. These scores are then combined to assign a score on the aforementioned 7-point scale of each dimension. An overview of all the indicators is provided in Table 3. The CLASS manual provides clear guidelines on how to weigh these indicator ratings in assigning a dimension score. For example, when all indicators are scored in the low range, the final score will

Table 2
Descriptive statistics of the CLASS Toddler dimensions.

CLASS Toddler dimension	*M*	*SD*	Range
Positive Climate	5.42	1.17	1–7
Negative Climate (reversed)	6.84	0.38	5–7
Teacher Sensitivity	5.34	1.08	2–7
Regard for Child Perspectives	4.24	1.34	1–7
Behavior Guidance	5.01	1.12	2–7
Facilitation of Learning and Development	3.73	1.35	1–7
Quality of Feedback	2.91	1.20	1–7
Language Modeling	3.22	1.29	1–7

$N = 1084$ number of observation cycles with the CLASS.

Table 3
Percentage of scores assigned at each rating category.

Dimensions	Indicator	Low (1)	Low/Mid(2)	Mid(3)	Mid/High(4)	High(5)
Positive Climate	Relationships	1%	4%	34%	24%	37%
	Positive Affect	1%	4%	34%	23%	37%
	Respect	1%	2%	30%	20%	47%
Negative Climate (reversed scale)	Negative affect	0%	0%	5%	3%	92%
	Punitive control	0%	0%	4%	3%	94%
	Teacher negativity	0%	0%	1%	1%	98%
	Child negativity	0%	0%	4%	5%	90%
Teacher Sensitivity	Awareness	1%	5%	42%	24%	29%
	Responsiveness	1%	3%	38%	23%	35%
	Child comfort	0%	2%	27%	23%	48%
Regard for Child Perspectives	Child focus	12%	15%	40%	14%	19%
	Flexibility	5%	11%	43%	14%	27%
	Support of independence	14%	13%	53%	11%	8%
Behavior Guidance	Proactive	2%	4%	41%	24%	29%
	Supporting positive behavior	5%	7%	48%	20%	21%
	Problem behavior	1%	4%	37%	20%	39%
Facilitation of Learning and Development	Active facilitation	26%	16%	37%	11%	10%
	Expansion of cognition	38%	19%	30%	8%	6%
	Children's active engagement	2%	11%	46%	14%	27%
Quality of Feedback	Scaffolding	41%	16%	35%	5%	3%
	Providing information	40%	14%	39%	4%	3%
	Encouragement and affirmation	28%	18%	44%	5%	6%
Language Modeling	Supporting language use	21%	17%	42%	8%	13%
	Repetition and extension	30%	17%	40%	6%	7%
	Self- and parallel talk	38%	11%	40%	5%	6%
	Advanced language	31%	20%	38%	6%	6%

consequently also be in the lowest range, that is, a 1; when all indicators are scored in the midrange, the final score would be exactly in the midrange, or a 4.

The dimension Positive Climate consists of three indicators. For example, the indicator positive affect evaluates whether teacher and children show enjoyment, enthusiasm and affection. Negative Climate consists of four indicators. An example is the indicator negative affect, which considers whether the teacher shows irritability, anger or uses a harsh voice in her interactions with children. Teacher Sensitivity consists of three indicators. For instance, the indicator child comfort assesses whether the children feel comfortable in approaching the teacher and seeking support or help from the teacher. Regard for Child Perspectives consists of three indicators. An example is the indicator child focus, which reflects the teacher's provision of choices for children, following their lead and eliciting children's ideas. Behavior Guidance consists of three indicators. For instance, the indicator proactive evaluates the teacher's use of monitoring and communication of clear behavioral expectations for the children. Facilitation of Learning and Development consists of three indicators. An example is the indicator children's active engagement, which reflects children's physical and verbal involvement in activities. Quality of Feedback consists of three indicators. For example, the indicator scaffolding captures the teacher's use of hints, verbal or physical assistance and prompting children's thought processes. Language Modeling consists of four indicators. For example, the indicator supporting language use captures the teacher's use of back-and-forth linguistic exchanges, contingent responding and open-ended questioning to elicit children's talk.

2.2.2. Self-reported developmental and educational activities

A structured questionnaire for teachers was used to assess the developmental and educational activities provided to the children on a regular basis over a longer period of time. This questionnaire was developed for the purposes of the current study, based on extant research on play, emergent literacy and emergent numeracy, and extensively tested in pilot research with teachers of two- and three-year-old children to ensure age-appropriateness of the listed activities (for more information, see Slot et al., 2015). Several scales were constructed covering a broad range of behaviors and activities. For the purpose of the current study, three types of activities were distinguished: play activities, literacy activities, and math activities.

The scale Play (9 items; $\alpha = 0.85$) assesses the degree to which the teacher provides children with opportunities for free, self-managed play and enriches their play, for instance by asking questions, making suggestions, or providing materials for richer play. Examples of items are: "I let the children play without interfering", "I ask children questions that stimulate their play". The scale ranges from 1 (*not applicable*) to 5 (*strongly applicable*).

The scale Literacy activities (4 items; $\alpha = 0.82$) measures the average frequency with which activities are provided involving literacy and literacy materials. An example of an item is: "Asking the children questions about the content of the story during or after reading the story". Answers were rated on a 7-point scale, ranging from 1 (*never*), 2 (*less than twice a month*), 3 (*twice or thrice a month*), 4 (*weekly*), 5 (*two to four times a week*), 6 (*daily*) to 7 (*three or more times a day*)

The scale Math activities (12 items; $\alpha = 0.91$) assesses the average frequency of several math activities, for instance counting and sorting activities, and activities exploring different shapes. An example of an item is: "Counting how many objects you have, for example, counting till five and saying 'I have five marbles'". Answers were rated on the same 7-point scale as the literacy activities scale.

2.2.3. Structural classroom and center characteristics

For the present purposes, the following structural quality variables were constructed using observations and teacher reported background information.

2.2.3.1. Education program. Teachers were asked whether they used an approved education program (see above). A dummy variable was created with the values 1 = *yes* and 0 = *no*, indicating whether an education program was used, without further distinctions between the programs.

2.2.3.2. Type of provision. The current study involved the two main types of ECEC in The Netherlands, namely daycare and preschool. A dummy variable was created with the values 1 = *preschool* and 0 = *daycare*.

2.2.3.3. Group size. CLASS observers recorded the number of children present during each observation cycle. An average score of the four observation cycles was used in subsequent analyses.

2.2.3.4. Children-to-teacher ratio. CLASS observers recorded the number of adults present during each observation cycle. The children-to-teacher ratio was calculated by dividing the number of children by the number of teachers. An average score based on the four observation cycles was used in the subsequent analyses.

2.2.3.5. Teacher's education. Teachers were asked to report their highest level of completed formal pre-service education on a scale representing the levels of Dutch secondary and tertiary education, ranging from 1 (*lower preparatory vocational track*) to 8 (*university*).

2.2.3.6. Work experience. Teachers were asked to report their work experience on a 7-point scale, ranging from 1 (*less than one year*) to 7 (>30 years).

2.2.3.7. Proportion of non-Dutch speaking children. Teachers were asked to report the proportion of children speaking another home language than Dutch on a 10-point scale, ranging from 1 (0–10%) to 10 (91–100%).

2.3. Procedures

Each classroom was observed with the CLASS Toddler during one morning. The observers rated classroom processes and teacher behavior during four 15 to 20 min cycles on the observation morning, resulting in a total of 1084 observation units. Classrooms were visited on a regular morning that was typical of the usual environment and routines (i.e., not during a day when a field trip was planned). Visits lasted three to four hours. The teacher questionnaire was sent out during the same period of data collection as the observations.

2.4. Analysis strategy

To address the three research aims several analyses were conducted, which are described in more detail below.

2.4.1. Structural validity

To address the first aim of this study, multilevel confirmatory factor analyses of the dimension and domain structure were conducted, taking the nested structure of the data into account and evaluating the equivalence of the factor structure at the within and between classrooms level using Mplus 7 (Muthén & Muthén, 1998–2012). The dimensional structure was tested using the CLASS Toddler indicators as observed variables. To account for the categorical measurement level of the indicators, model building included two assumptions originating in IRT: (1) the observed classroom processes can be ordered per indicator on an ordinal scale according to their level of quality; (2) there is an underlying continuous latent ability, that is, the teacher's ability to provide a particular level of process quality in the classroom, that predicts the probabilities with which each distinguished level of quality will be observed (de Ayala, 2013). Next, the domain structure was tested using the CLASS Toddler dimensions as observed variables in CFAs following a CTT approach.

2.4.2. Item properties of the CLASS Toddler indicators

Addressing the second aim, the CLASS Toddler indicators were examined with an IRT approach, following the best fitting domain structure of the previous analysis. Two common item statistics were computed: item difficulty (by calculating the mean of all thresholds per item) and item discrimination (represented by the standardized factor loadings). Item difficulty is the average difficulty level that locates the indicator along the latent quality scale. Difficulty estimates are represented on a Logit scale, with the mean arbitrarily set to zero and with lower (negative) values indicating easier items (most teachers are most likely to be rated in the higher score categories), values around 0 indicating average difficulty and higher (positive) values indicating harder items (de Ayala, 2013). Discrimination estimates, expressed as standard scores with values between 0 and 1, are indicative of the relative amount of information provided by each indicator. A value of at least 0.30 is considered to indicate sufficient discrimination power and was used as the cut-off (Abell, Springer, & Kamata, 2009).

2.4.3. Criterion validity

To address the third aim, the relations between the best fitting domain model of the CLASS Toddler and report-based teacher, classroom, and curriculum characteristics were examined. For continuous variables, Pearson correlations were computed; for dichotomous variables, *t*-tests were conducted. Due to the design of the study and non-response, complete observational and self-report data were available for 39.8% of the classrooms for this part of the analysis. Importantly, classrooms with and without self-reports did not differ significantly on any of the CLASS Toddler dimensions, except that classrooms with observations had slightly lower scores on self-reported math activities than classrooms for which no observations were available (Cohen's $d = 0.10$). There were no further missing data in the observation measures. Occasionally missing data on activities and structural classroom characteristics in the self-reports were below 8%. Missing data were dealt with by using full information maximum likelihood (FIML) estimation in Mplus (Enders, 2010).

2.4.4. Evaluation model fit

All IRT-based analyses were run using a means and variances adjusted weighted least squares (WLSMV) estimator, which yields good approximations of parameter estimates with ordinal data (Hill et al., 2007). As only the RMSEA fit index has proven to be useful in evaluating the model fit of IRT models (Maydeu-Olivares, Cai, & Hernandez, 2011), RMSEA < 0.05 was considered to indicate good fit and RMSEA < 0.08 to indicate acceptable fit. No other fit indices were used. Model fit in CTT-based analyses was evaluated following the usual criteria, with Chi-Square/df < 3, CFI/TLI > 0.95, and RMSEA, $SRMR_{within}$ and $SRMR_{between}$ < 0.05, indicating good fit, and CFI/TLI > 0.90 and RMSEA, $SRMR_{within}$ and $SRMR_{between}$ < 0.08 indicating acceptable fit. The Chi-Square/df ratio rather than the Chi-Square test was used, as this fit index is best suited for large samples (Kline, 2005). Improvement of the model fit was evaluated through Chi-Square difference tests.

3. Results

Table 3 shows the distribution of the scores over the five score categories for each indicator of the eight dimensions. Overall, the score distributions show a reasonably ordered pattern, with clear peaks in observed frequencies indicating the predominant level of quality as measured by the particular indicator. Table 3 also shows that indicators differ in score distribution, with sometimes the peak in the lower score range and sometimes in the higher score range. Notably, the indicators of Negative Climate (reversed scale) reveal hardly any variation, with all scores peaking in the mid-to-high range, indicating that negativity was rare in the observed classrooms. Furthermore, although all score levels were observed, most scores were in the mid-to-high range for the indicators of Positive Climate, Teacher Sensitivity, Regard for Child Perspectives and Behavior Guidance. The indicators of Facilitation of Learning and Development, Quality of Feedback, and Language Modeling showed the reversed pattern with most scores in the low-to-mid range and few scores in the mid-to-high range.

3.1. Structural validity

First, a multilevel confirmatory factor analysis for categorical data was conducted on all indicators to test the eight-dimension structure

Table 4
Factor loadings at the within and between level for all indicators in the Eight-Factor Model (based on IRT analyses).

Dimensions	Indicator	Factor loading within level	Factor loading between level
Positive Climate	Relationships	0.57	0.77
	Positive Affect	0.64	0.85
	Respect	0.56	0.99
Negative Climate	Negative affect	0.65	0.85
	Punitive control	0.47	0.46
	Teacher negativity	0.21	1.00[1]
	Child negativity	0.58	0.57
Teacher Sensitivity	Awareness	0.61	0.71
	Responsiveness	0.60	0.90
	Child comfort	0.52	0.88
Regard for Child Perspectives	Child focus	0.83	0.84
	Flexibility	0.77	0.73
	Support of independence	0.48	0.69
Behavior Guidance	Proactive	0.58	0.80
	Supporting positive behavior	0.53	0.65
	Problem behavior	0.60	0.74
Facilitation of Learning and Development	Active facilitation	0.75	0.91
	Expansion of cognition	0.81	0.85
	Children's active engagement	0.51	0.76
Quality of Feedback	Scaffolding	0.54	0.54
	Providing information	0.72	0.84
	Encouragement and affirmation	0.51	0.77
Language Modeling	Supporting language use	0.75	0.80
	Repetition and extension	0.61	0.79
	Self- and parallel talk	0.36	0.52
	Advanced language	0.68	0.65

[1] Value after constraining the residual variance to 0.001.

of the CLASS Toddler using an IRT approach. Details of the model are given in Table 4.

The RMSEA was 0.04, which indicates overall good fit. As can be seen in Table 4, most factor loadings were acceptable to good for all dimensions, particularly at the between classroom level. Regarding the dimension Negative Climate, the indicator teacher negativity, however, was problematic. The variance was very limited, as is also evident from Table 3, and could not be reliably estimated in Mplus, resulting in a low non-significant factor loading at the within level and a high factor loading exceeding 1 at the between level. The problem remained after fixing the residual variance to 0.001, indicating that this factor loading could not be readily interpreted. In addition, the correlations presented in Table 5 showed a strong correlation between Negative Climate and Behavior Guidance, indicating that these dimensions may actually represent a single factor.

Next, the proposed two-factor model with Emotional and Behavioral Support and Engaged Support for Learning as overarching domains was evaluated using a multilevel confirmatory factor analysis with the eight CLASS Toddler dimensions as observed variables, following a CTT approach. Model testing revealed overall good fit and a reasonably equivalent factor structure at the within and between level (see Table 6). However, further inspection of the standardized factor loadings showed a relatively low factor loading for Negative Climate at both the within level (0.35) and the between level (0.45) on the Emotional and Behavioral Support domain (see Table 6). Moreover, the dimension Regard for Child Perspectives showed a low non-significant factor loading on the domain Emotional Support at the within level (0.12), but a high factor loading at the between level (0.89).

The two-factor model was tested against two alternative models. The first was a one-factor model representing general classroom process quality (in line with the so called Effective teaching model; Hamre et al., 2013), which showed poor overall model fit and a significant deterioration of the model compared to the two-factor model, ($\Delta\chi^2(2) = 310.02$, $p < 0.001$). Also, the factor loadings in this model were generally lower for all dimensions at both the within and between level, particularly for the dimensions Negative Climate and Behavior Guidance. The second alternative model was a multilevel three-factor model. Based on the inter-correlations presented in the previous section, we distinguished the domains Emotional Support (indicated by the dimensions Positive Climate, Teacher Sensitivity, and Regard for Child Perspectives), Behavioral Support (indicated by Negative Climate and Behavior Guidance), and Engaged Support for Learning (indicated by Facilitation of Learning and Development, Quality of Feedback, and Language Modeling). The three-factor model showed good model fit ($\chi^2(34) = 82.50$, $p < 0.001$; $\chi^2/df = 2.43$; RMSEA = 0.04; CFI = 0.98, TLI = 0.96; $SRMR_{within/between} = 0.03/0.04$) and significantly improved model fit compared to the two-factor model ($\Delta\chi^2(4) = 19.07$ $p < 0.001$). In addition, the factor loadings of both Negative Climate and Behavior Guidance were higher in the three-factor model compared to the two-factor model, both at the within and at the between classrooms level. Note, however, that the factor loading of Regard for Child Perspectives at the within level remained small. We will return to this issue in the Discussion. To conclude, the three-factor model, representing three domains of classroom process quality, showed the best fit to the data, accounted for more observed variance, and revealed a reasonably equivalent factor structure at both the level of the observation cycles and the classroom level.

3.2. Item properties of the CLASS Toddler indicators

Following the previous analyses in which the three-domain model showed the best model fit and a sufficiently equivalent factor structure within and between classrooms, we reexamined the final model with an IRT approach to obtain estimates of the item difficulty and item discrimination parameters for all indicators, as shown in Table 7. The results for item difficulty showed that almost all indicators of Positive Climate, Negative Climate, Teacher Sensitivity, Regard for Child Perspectives, and Behavior Guidance were negative or close to zero, indicating that teachers were likely to have high scores on these dimensions. The

Table 5
Between-level intercorrelations among CLASS dimensions in the eight-factor model (based on IRT analyses).

CLASS dimensions		1	2	3	4	5	6	7	8
1	Positive Climate		0.56***	0.83***	0.79***	0.68***	0.81***	0.83***	0.92****
2	Negative Climate			0.59***	0.64***	0.76***	0.41***	0.61***	0.51***
3	Teacher Sensitivity				1.04***	0.96***	0.76***	0.73***	0.80***
4	Regard Child Pers.					0.95***	0.97***	1.01***	0.90***
5	Behavior Guidance						0.61***	0.63***	0.61***
6	Facilitation Learning							1.04***	1.06***
7	Quality of Feedback								1.10***
8	Language Modeling								

*** $p < 0.001$.

Table 6
Fit indices for the multilevel one-factor, two-factor, and three-factor models (based on CTT).

Factor loadings at the within (W) and between (B) level	Two-factor model		One-factor model		Three-factor model	
	W	B	W	B	W	B
Positive Climate	0.66	0.81[1]	0.49	0.64	0.67	0.80[1]
Negative Climate	0.35	0.45[1]	0.09	0.09	0.37	0.51[2]
Teacher Sensitivity	0.69	0.90[1]	0.49	0.51	0.70	0.89[1]
Regard for Child Perspectives	0.12	0.89[1]	0.17	0.53	0.13	0.89[1]
Behavior Guidance	0.59	0.78[1]	0.46	0.43	0.65	0.94[2]
Facilitation of Learning and Development	0.70	0.95[2]	0.71	0.72	0.71	0.95[3]
Quality of Feedback	0.67	0.89[2]	0.57	0.65	0.67	0.88[3]
Language Modeling	0.68	0.95[2]	0.53	0.88	0.68	0.95[3]
Measures of fit information						
Chi-square value	101.57		411.59		82.50	
df	38		40		34	
Ratio Chi-square/*df*	2.67		10.29		2.43	
RMSEA	0.04		0.09		0.04	
CFI	0.97		0.83		0.98	
TLI	0.96		0.77		0.96	
$SRMR_{within}$	0.03		0.06		0.03	
$SRMR_{between}$	0.05		0.09		0.04	

Note: Values with the same superscript numbers belong to the same factor.

indicators of the dimension Negative Climate (reversed scale) showed the lowest item difficulty indicating that teachers were likely to receive a high score on this dimension, which is in accordance with the fact that behaviors indicating a Negative Climate were hardly observed. The dimensions Facilitation of Learning and Development, Quality of Feedback, and Language Modeling revealed a reversed pattern, indicating that teachers were less likely to have high scores on the indicators underlying these dimensions. Discrimination parameters of all indicators ranged from 0.412 to 0.803, indicating that all indicators showed moderate to good discrimination with values above the recommended cut-off of 0.30 (Abell et al., 2009). To conclude, the set of indicators that was used to obtain the primary behavioral data for the ECEC quality assessment with the CLASS Toddler met the requirements of good measurement quality adequately.

3.3. Criterion validity

Table 8 shows the correlations between the three CLASS Toddler domains and several teacher and classroom characteristics at the classroom level. There were a number of significant but small-sized correlations between the domains and the structural quality characteristics, which were all in the expected direction. A higher children-to-teacher ratio was related to both lower Emotional Support and Engaged Support for Learning, but not to Behavioral Support. Teacher's qualifications and group size were not related to the CLASS Toddler domains, nor was working with an education program. Teachers' work experience was positively related to Engaged Support for Learning, but not to the two other domains. Furthermore, in classrooms with a higher proportion of non-Dutch speaking children, teachers showed higher Engaged Support for Learning. Preschools had significantly higher levels of Engaged Support for Learning than daycare centers. In addition, there were several expected, weak to moderate correlations between observed classroom quality and the reported curriculum activities, particularly for the Engaged Support for Learning domain. The provision of literacy activities and, to a lesser extent, the provision of play was positively related to observed process quality. No correlations were found with the provision of math activities. To conclude, the expected weak to moderate associations between the CLASS Toddler domains and several structural quality and curriculum characteristics indicated adequate criterion validity.

4. Discussion

The CLASS Toddler, like other well-known observational measures of ECEC process quality, is increasingly used for scientific as well as accountability and professionalization purposes with possibly important implications for policy and practice. However, research on the reliability and validity of the CLASS Toddler is still scarce and limited to US contexts. The current study aimed to fill part of this gap by investigating the measurement properties of the CLASS Toddler in Dutch ECEC. We combined a Classical Test Theory (CTT) with an Item Response Theory (IRT) approach to evaluate the CLASS Toddler at the indicator, dimension and domain level. Moreover, we used a multilevel design to take

Table 7
Overall item difficulty and item discrimination based on the three-factor model.

Domain	Dimensions	Indicator	Difficulty	Discrimination
Emotional Support	Positive Climate	Relationships	−0.890	0.703
		Positive affect	−0.861	0.800
		Respect	−1.093	0.802
	Teacher Sensitivity	Awareness	−0.692	0.654
		Responsiveness	−0.840	0.775
		Child comfort	−1.160	0.737
	Regard for Child Perspectives	Child focus	−0.116	0.755
		Flexibility	−0.449	0.877
		Support of independence	0.140	0.610
Behavioral Support	Negative Climate (reversely coded)	Negative affect	−1.526	0.905
		Punitive control	−1.667	0.626
		Teacher negativity	−2.236	0.659
		Child negativity	−1.523	0.699
	Behavior Guidance	Proactive	−0.692	0.693
		Supporting positive behavior	−0.417	0.610
		Problem behavior	−0.885	0.651
Engaged Support for Learning	Facilitation of Learning and Development	Active facilitation	0.338	0.751
		Expansion of cognition	0.676	0.795
		Children's active engagement	−0.495	0.527
	Quality of Feedback	Scaffolding	0.833	0.504
		Providing information	0.788	0.734
		Encouragement and affirmation	0.536	0.591
	Language Modeling	Supporting language use	0.235	0.821
		Repetition and extension	0.537	0.707
		Self- and parallel talk	0.670	0.513
		Advanced language	0.590	0.704

Table 8
Associations between CLASS domains and teacher and classroom characteristics.

	Emotional support	Behavioral support	Engaged support for learning
Categorical classroom characteristics and CLASS domains and *t*-tests between subgroups			
Education program			
Yes	4.94	5.92	3.25
No	5.02	5.78	3.23
Type of provision			
Day care	5.04	5.88	3.08 *
Preschool	4.96	5.94	3.44 *
Continuous teacher and classroom characteristics and Pearson correlations			
Pre-service education level	−0.03	0.05	−0.01
Group size	−0.06	−0.02	−0.03
Children-to-teacher ratio	−0.12†	−0.10	−0.16**
Work experience	0.14	0.16	0.19*
% of non-Dutch speaking children	0.03	−0.04	0.16†
Provision of activities (based on self-reports) and Pearson correlations			
Play	0.19*	0.07	0.27*
Literacy	0.17†	0.22*	0.25**
Math	−0.05	−0.05	0.06

* $p < 0.05$.
** $p < 0.01$.
† $p < 0.10$.

the nested structure of the recommended cyclic observations into account. The following three aspects of the CLASS Toddler were investigated: (i) structural validity, (ii) measurement properties of the indicators, and (iii) criterion validity.

4.1. Structural validity

Although a two-domain structure as proposed by the developers of the CLASS Toddler (La Paro et al., 2014) showed good model fit, the results of multilevel factor analyses indicated that a three-domain structure provided the best fit to our data, with Behavioral Support (consisting of the dimensions Negative Climate and Behavior Guidance) constituting a separate third domain. Other non-US studies revealed patterns of correlations between these dimensions that are largely in line with the current findings (Leyva et al., 2015; Pakarinen et al., 2010; von Suchodoletz et al., 2014). Besides better fit, the three-domain model also showed higher factor loadings for the Negative Climate and Behavior Guidance dimensions, thus accounting for more observed variance. The present results contrast with findings in US studies in which a two-domain structure was found to be sufficient to capture the relevant variance (Bandel et al., 2014; Castle et al., 2016). The present findings thus may point to cultural differences between the US, on the one hand, and The Netherlands and several other countries, on the other hand.

Although the scores on the dimension Negative Climate are generally low across countries, they appeared to be higher for US classrooms compared to The Netherlands. For example, compared to a mean score of 1.16 in The Netherlands on this dimension, with a small standard deviation of 0.38, the mean scores in a number of US studies ranged from 1.30 to 2.70 depending on the sample and the type of program, with much larger standard deviations ranging from 0.71 to 1.29 (La Paro et al., 2014; Thomason & La Paro, 2009; Vogel, Caronongan, Xue et al., 2015). In the Dutch context, teachers hardly ever showed humiliating, teasing or sarcastic behavior while interacting with children. The Negative Climate dimension, therefore, reflected only mild teacher negativity, such as occasionally raising the voice and expressing irritation in relation to specific events in the classroom, including incidental negative behavior of children.

Behavior Guidance was scored slightly higher in Dutch classrooms compared to findings in some US studies with substantial differences of about one scale point (La Paro et al., 2014; Thomason & La Paro, 2009). In this regard, the relationship between Negative Climate and Behavior Guidance could be understood as indicating that effective behavioral support, through providing clear behavioral expectations and supporting and reinforcing children's positive behavior, is associated with less negativity in the classroom, whereas less effective behavioral support is associated with higher negativity. Support for this hypothesis comes from cross-cultural research showing a typical model of caring for children among Dutch parents as well as professional caregivers marked by emphasis on regularity (Harkness & Super, 2006, p. 68) and by the use of fixed time schedules (Slot et al., 2015) that in ECEC results in a rather predictable environment in which children know what is happening next, allowing them to adapt their behavior accordingly.

The multilevel analyses of the dimensional structure of the CLASS Toddler showed that the factor structures at the within and between classrooms level were sufficiently equivalent, indicating that the three domains represented largely equivalent quality constructs across levels. The only exception was the dimension Regard for Child Perspectives that appeared to fit less well in the domain Emotional Support at the within level. These findings confirm that the dimension Regard for Child Perspectives is part of the Emotional Support domain, as proposed by the developers of the CLASS Toddler, when the cycle scores are aggregated to the classroom level. However, our findings also suggest that this domain represents a different construct when used to assess and compare the quality of different situations within classrooms. Put differently, the quality construct used to evaluate situations within classrooms differs slightly in meaning from the quality construct used to evaluate and compare classrooms. This finding may point to a cultural difference as well.

In order to explore this possible explanation, we examined our data more closely and found a typical pattern of scores across situations on the dimensions constituting the Emotional Support domain (see Slot et al., 2015). In play situations, Regard for Child Perspectives was relatively high ($M = 5.15$, $SD = 1.06$) due to the fact that children were allowed a lot of freedom and initiative (play was mostly free, unguided play), whereas in the other observed situations such as teacher guided educational, creative and care activities, the score for Regard for Child Perspectives was relatively low ($M = 3.83$, $SD = 0.85$). In contrast, the scores on the other dimensions of the Emotional Support domain, especially Positive Climate and Teacher Sensitivity, were relatively low during play, due to low teacher involvement, and high in the other situations. Put differently, in the Dutch context, Regard for Child Perspectives, when averaged over different situations (between classroom level), was clearly related to the other dimensions of the Emotional Support domain, as was expected on the basis of US research. However, the main way in which Dutch teachers allowed children initiative and freedom of choice was by providing them with free play during particular time slots on a day in which a high degree of child initiative and freedom of choice was observed (but also lower teacher involvement, less teacher sensitivity and more negativity among children), whereas all other situations were more strongly teacher-directed (thus lower on Regard for Child Perspectives, but, as was found, higher on the other Emotional Support dimensions). To what extent this pattern of implementing Regard for Child Perspectives is typical for Dutch ECEC or is also present in other cultural contexts is not clear, as no previous studies on ECEC classroom quality have compared factor models of process quality at both the within and between level. Note that the present findings underscore that a multilevel approach can be informative by revealing that the same aggregated process quality score at the classroom level can represent different models of implementing process quality within classrooms.

4.2. Item properties of the CLASS Toddler indicators

The set of CLASS Toddler indicators was found to be adequate overall, showing the desired variation in item difficulty and sufficient discrimination values for all indicators. Although teachers were more

likely to receive high scores on the indicators of the domains Emotional Support and Behavioral Support than on the indicators of the domain Engaged Support for Learning, there was still considerable variance. Moreover, the discrimination values showed that the set of indicators for each quality dimension enabled to distinguish well between teachers in both the low and high range of this dimension. Interestingly, the indicators of the three domains Emotional Support, Behavioral Support, and Engaged Support for Learning had equally high discrimination values, indicating that the differences in item difficulties (the indicators of the domains Emotional Support and Behavioral Support being 'easier' than the indicators of the domain Engaged Support for Learning) most likely reflect that Dutch ECEC teachers are better able to provide high emotional and behavioral process quality than high instructional process quality. Note that, despite the limited variability in Negative Climate, all indicators of this dimension showed satisfactory discrimination values and thus contributed relevant information to the model.

4.3. Criterion validity

Finally, support was found for the criterion validity of the CLASS Toddler. Structural aspects, such as teacher-child ratio and teachers' work experience were related to all process quality domains, in line with previous findings on the CLASS Toddler (Slot et al., 2015; Thomason & La Paro, 2009), as well as with findings in Dutch studies using similar observation instruments for evaluating process quality (de Kruif et al., 2009; Fukkink et al., 2013; Vermeer et al., 2008). Teachers' qualifications and group size were not related to the CLASS domains, in accordance with similar findings in previous research (Castle et al., 2016; Howes et al., 2008; LoCasale-Crouch et al., 2007; Mashburn et al., 2008; Pianta et al., 2005; Vogel, Caronongan, Thomas et al., 2015; Vogel, Caronongan, Xue et al., 2015). A possible explanation is the restricted range of variance due to statutory quality regulations (Castle et al., 2016; Love et al., 2003). This holds especially for the aforementioned structural quality aspects of Dutch ECEC (de Kruif et al., 2009; Fukkink et al., 2013; Slot et al., 2015; Vermeer et al., 2008). Preschools provided higher support for children's learning, which is in line with their stronger educational orientation and corroborates previous findings (Slot et al., 2015). However, the use of an education program was unexpectedly not related to the observation measures, also not to Engaged Support for Learning, which may point to low implementation quality of such programs in The Netherlands (Doolaard & Leseman, 2008).

Furthermore, several correlations between the CLASS domains and teachers' self-reported curriculum activities were found. The reported enrichment of children's play was related to observed Emotional Support, but especially to observed Engaged Support for Learning. This finding is in line with the results of a previous study from the US (Cabell et al., 2013), but contradicts recent findings from a study on daycare in The Netherlands (Singer et al., 2014). A possible explanation for this discrepancy is the emphasis on teachers' stimulation of play in the current study, thus assuming an active role of teachers, whereas Singer et al. (2014) observed mainly free play with limited teacher involvement. The reported provision of literacy activities correlated especially with Engaged Support for Learning, as was expected (Cabell et al., 2013). However, for math activities no such relation was found, unlike in previous US research on Head Start preschools (Cabell et al., 2013). A possible explanation for this discrepancy in results is the rather infrequent provision of math activities in Dutch ECEC (De Haan et al., 2013; Slot et al., 2015).

4.4. Limitations

There are a number of limitations to the present study. A first limitation is that the current study did not examine to what extent scores for an ECEC center can be generalized beyond the measurement occasion. Future studies should evaluate score generalizability by examining stability of scores across measurement occasions. Related to this, future studies should also evaluate agreement between observers during data collection. In the current study, all observers conducted a live observation with a licensed CLASS Toddler trainer in which they showed high inter-rater agreement, but this was only prior to data collection. A second limitation is that no data on children's development were included to relate the quality measures to child outcomes as indication of predictive validity. Although such relations have been reported in several previous studies (Bandel et al., 2014; La Paro et al., 2014), it is nonetheless important to further validate the CLASS Toddler by including children's socio-emotional and cognitive outcomes. A third limitation concerns the evaluation of the criterion validity of the CLASS. Due to missing data, there was only limited overlap between the CLASS Toddler observations and the teacher reports, reducing the sample size for the criterion validity analysis. However, sufficient power remained and no systematic differences between centers with and without self-reports were found. Furthermore, the correlations of the three CLASS Toddler domains with structural quality and curriculum characteristics were small to moderate. Although our results are in line with previous Dutch research, stronger evidence concerning criterion validity is desirable.

5. Conclusion

The present study replicated previous studies into the measurement quality of the CLASS in a different national context, but differed in two respects from these earlier studies. First, the measurement properties of the CLASS Toddler indicators were examined. Second, a multilevel approach was applied to deal with the within-classroom variance in observed process quality. Despite these differences between the current study and earlier work, the present findings largely converge with previous findings from the US and other countries. This convergence can be explained by the fact that the set of indicators had good measurement properties and confirmed the proposed dimensional structure of the CLASS Toddler. Moreover, the factor structures at the within and between level proved to be sufficiently equivalent, meaning that a CFA with only aggregated classroom level dimension scores (as applied in the previous studies) should reveal the same results.

Despite its overall convergence with previous findings, the present study also found indications of possible cultural differences. The results regarding the dimension Negative Climate pointed to different interaction and classroom management styles on part of the teachers, and favored a three-domain structure over a two-domain structure, with Behavioral Support as an additional domain. Moreover, although sufficient overall equivalence of the factor structures at the within and between level was found, a discrepancy between the levels regarding the dimension Regard for Child perspectives, could be related to possible cultural differences in classroom practices as well. To conclude, despite these issues, the present study revealed good measurement properties of the CLASS Toddler.

With respect to the discrepant findings with previous US research in particular, it is recommended to re-analyze the CLASS Toddler as representing three domains of process quality, also in view of the fact that several other non-US studies show similar outcomes as the present study. This would strengthen the coherence in ECEC process quality measures across studies and countries. Preferably, future studies should investigate the measurement properties of the CLASS Toddler and other versions of the CLASS at the indicator, dimension and domain level, while taking both the within and between classrooms variance into account, as in the present study.

Acknowledgement

The research reported in this article was carried out with a grant of the Dutch National Science Foundation (NWO), project number 411-20-442.

References

Abell, N., Springer, D. W., & Kamata, A. (2009). *Developing and validating rapid assessments instruments.* New York, NY: Oxford University Press.

Ainsworth, M. D., Blehar, M. C., Waters, E., & Wall, D. (1978). *Patterns of attachment: A psychological study of the strange situation.* Hillsdale, NJ: Erlbaum.

de Ayala, R. J. (2013). *Theory and practice of item response theory (pp.1–38, 99–122).* New York, NY: Guilford Press.

Bandel, E., Aikens, N., Vogel, C. A., Boller, K., & Murphy, L. (2014). Observed quality and psychometric properties of the CLASS-T in the early head starts family and child experiences survey. *OPRE Technical Brief 2014-34.* Washington D.C.: Office of Planning, Research and Evaluation, Administration for Children and Families, US Department of Health and Human Services.

Blair, C. (2003). Behavioral inhibition and behavioral activation in young children: Relations with self-regulation and adaptation to preschool in children attending head start. *Developmental Psychobiology*, *42*(3), 301–311.

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

Brancheorganisatie Kinderopvang (2014). Brancherapport Kinderopvang. Kinderopvang 2014: feiten, cijfers & ontwikkelingen [Report Child care 2014: facts, numbers & developments]. Retrieved from http://www.kinderopvang.nl/publicaties/onderzoeken (on 3-19-2015)

Bronfenbrenner, U., & Morris, P. A. (2006). The bioecological model of human development. In R. M. Lerner (Ed.), *Theoretical models of human development, Handbook of Child Psychology (6th ed.)vol. 1.* (pp. 793–828). Hoboken, NJ: Wiley.

Bronson, M. B. (2000). *Self-regulation in early childhood. Nature and nurture.* New York, NY: The Guilford Press.

Bryant, D. M., Burchinal, M., & Zaslow, M. (2010). Empirical approaches to strengthening the measurement of quality: Issues in the development and use of quality measures in research and applied settings. In M. Zaslow (Ed.), *Quality measurement in early childhood settings* (pp. 33–51). Baltimore, MD: Paul H. Brookes.

Burchinal, M., Kainz, K., & Cai, Y. (2011). How well do our measures of quality predict child outcomes? A meta-analysis and coordinated analysis of data from large-scale studies of early childhood settings. In M. Zaslow (Ed.), *Quality measurement in early childhood settings* (pp. 11–33). Baltimore, MD: Paul H. Brookes.

Cabell, S. Q., DeCoster, J., LoCasale-Crouch, J., Hamre, B. K., & Pianta, R. C. (2013). Variation in the effectiveness of instructional interactions across preschool classroom settings and learning activities. *Early Childhood Research Quarterly*, *28*, 820–830. http://dx.doi.org/10.1016/j.ecresq.2013.07.007.

Cadima, J., Leal, T., & Burchinal, M. (2010). The quality of teacher-student interactions: Associations with first graders' academic and behavioral outcomes. *Journal of School Psychology*, *48*, 457–482. http://dx.doi.org/10.1016/j.jsp.2010.09.001.

Cassidy, D. J., Hestenes, L. L., Hegde, A., Hestenes, S., & Mims, S. (2005). Measurement of quality in preschool child care classrooms: An exploratory and confirmatory factor analysis of the early childhood environment rating scale-revised. *Early Childhood Research Quarterly*, *20*, 345–360. http://dx.doi.org/10.1016/j.ecresq.2005.07.005.

Castle, S., Williamson, A. C., Young, E., Stubblefield, J., Laurin, D., & Pearce, N. (2016). Teacher-child interactions in early had start classrooms: Associations with teacher characteristics. *Early Education and Development*, *27*, 259–274. http://dx.doi.org/10.1080/10409289.2016.1102017.

Catts, H. W., Fey, M. E., Zhang, X., & Tomblin, J. B. (1999). Language basis of reading and reading disabilities: Evidence from a longitudinal investigation. *Scientific Studies of Reading*, *3*(4), 331–361.

Colwell, N., Gordon, R. A., Fujimoto, K., Kaestner, R., & Korenman, S. (2013). New evidence on the validity of the Arnett caregiver interaction scale: Results for the early childhood longitudinal study-birth cohort. *Early Childhood Research Quarterly*, *28*, 218–233. http://dx.doi.org/10.1016/j.ecresq.2012.004.

Convenant Kwaliteit Kinderopvang. Bijlage bij Nieuw Convenant Kinderopvang [Convenant Quality Child Care. Appendix to New Convenant Child Care]. Retrieved from http://www.rijksoverheid.nl/onderwerpen/kinderopvang/documenten-en-publicaties/kamerstukken/2008/11/02/bijlage-convenant-kwaliteit-kinderopvang.html(2008).

Cryer, D., Tietze, W., Burchinal, M., Leal, T., & Palacios, J. (1999). Predicting process quality from structural quality in preschool programs: A cross-country comparison. *Early Childhood Research Quarterly*, *14*, 339–361.

Davis, E. A., & Miyake, N. (2004). Explorations of scaffolding in complex classroom systems. *Journal of the Learning Sciences*, *13*(3), 265–272.

Doolaard, S., & Leseman, P. P. M. (2008). *Versterking van het fundament. Integrerende studie n.a.v. de opbrengsten van de onderzoekslijn sociale en institutionele context van scholen uit het onderzoeksprogramma beleidsgericht onderzoek primair onderwijs 2005–2008 [Strengthening the foundation. Integrating study following the returns of the researchline social and institutional context of schools from the Research Programme Primary Education 2005–2008].* Groningen: GION.

De Haan, A., Elbers, E., Hoofs, H., & Leseman, P. (2013). Targeted versus mixed preschool and kindergartners: effects of classroom composition and teacher-managed activities on disadvantaged children's emergent academic skills. *School Effectiveness and School Improvement: An International Journal of Research, Policy, and Practice*, *24*, 177–194. http://dx.doi.org/10.1080/0924353.2012.74792.

Emmer, E. T., & Strough, L. (2001). Classroom management: A critical part of educational psychology, with implications for teacher education. *Educational Psychologist*, *36*(2), 103–112.

Enders, C. K. (2010). *Applied missing data analysis.* New York, NY: Guilford Press.

Fukkink, R. G., Gever Deynoot-Schaub, M. J. J. M., Helmerhorst, K. O. W., Bollen, I., & Riksen-Walraven, J. M. A. (2013). *Pedagogische kwaliteit van de kinderopvang voor 0–4 jarigen in Nederlandse kinderdagverblijven in 2012 [Pedagogical quality of Dutch child care for 0- to 4- years-olds in Dutch daycare centers in 2012].* Amsterdam/Nijmegen: NCKO.

Gordon, R. A., Fujimoto, K., Kaestner, R., Korenman, S., & Abner, K. (2013). An assessment of the validity of the ECERS-R with implications for measures of child care quality an relations for child development. *Developmental Psychology*, *49*, 146–160. http://dx.doi.org/10.1037/a002789.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38–47.

Hamre, B. K., & Pianta, R. C. (2007). Learning opportunities in preschool and early elementary classrooms. In R. C. Pianta, M. J. Cox, & K. Snow (Eds.), *School readiness, early learning and the transition to kindergarten* (pp. 49–84). Baltimore: Brookes Publishing.

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

Harkness, S., & Super, C. M. (2006). *Themes and variations: Parental ethnotheories in Western cultures. Parenting beliefs, behaviors, and parent-child relations: A cross-cultural perspective, 61–79.*

Helmerhorst, K. O. W., Riksen-Walraven, M. J., Vermeer, H. J., Fukkink, R. G., & Tavecchio, L. W. C. (2014). Measuring the interactive skills of caregivers in child care centers: Development and validation of the Caregiver Interaction Profile Scales. *Early Education and Development*, *25*, 1–21. http://dx.doi.org/10.1080/10409289.2014.840482.

Hestenes, L. L., Kintner-Duffy, V., Chen Wang, Y., La Paro, K., Mims, S. U., Crosby, D., ... Cassidy, D. J. (2015). Comparisons among quality measures in child care settings: Understanding the use of multiple measures in North Carolina's QRIS and their links to social-emotional development in preschool children. *Early Childhood Research Quarterly*, *30*, 199–214. http://dx.doi.org/10.1016/j.ecresq.2014.06.003.

Hill, C. D., Edwards, M. C., Thissen, D., Langer, M. M., Wirth, M. A., ... Varni, J. W. (2007). Practical issues in the application of item response theory. *Medical Care*, *45*, 39–47.

Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R., & Barbarin, O. (2008). Ready to learn? Children's pre-academic achievement in pre-Kindergarten programs. *Early Childhood Research Quarterly*, *23*, 27–50. http://dx.doi.org/10.1016/j.ecresq.2007.05.002.

Hox, J. J. (2010). *Multilevel analysis. Techniques and applications* (2nd ed.). New York: Routledge.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.

Kopp, C. B. (1982). Antecedents of self-regulation: a developmental perspective. *Developmental Psychology*, *18*, 199–214.

De Kruif, R. E. L., Riksen-Walraven, J. M., Gevers Deynoot-Schaub, M. J. J. M., Helmerhorst, K. O. W., Tavecchio, L. W. C., & Fukkink, R. G. (2009). *Pedagogogische kwaliteit van de opvang voor 0- tot 4-jarigen in Nederlandse kinderdagverblijven in 2008 [Pedagogical quality of Dutch child care for 0- to 4-years-olds in 2008].* Amsterdam/Nijmegen: NCKO.

De Kruif, R. E. L., McWilliam, R. A., Maher Ridley, S., & Wakely, M. B. (2000). Classification of teachers' interaction behaviors in early childhood classroom. *Early Childhood Research Quarterly*, *15*, 247–268. http://dx.doi.org/10.1016/S0885-2006(00)00051-X.

Lambert, R. G., Nelson, L., Brewer, D., & Burchinal, M. R. (2006). Measurement issues and psychometric methods in developmental research. *Monographs of the Society for Research in Child Development*, *71*(3), 24–41. http://dx.doi.org/10.1111/j.1540-5834.2006.00403.x.

La Paro, K. M., Hamre, B. K., & Pianta, R. C. (2011). *Classroom assessment scoring system toddler manual.* Charlottesville, VA: Teachstone.

La Paro, K. M., Williamson, A. C., & Hatfield, B. (2014). Assessing quality in toddler classrooms using the CLASS-Toddler and the ITERS-R. *Early Education and Development*, *25*, 875–893. http://dx.doi.org/10.1080/10409289.2014.883586.

Layzer, J. I., & Goodson, B. D. (2006). The "quality" of early care and education settings: Definitional and measurement issues. *Evaluation Review*, *30*, 556–576. http://dx.doi.org/10.1177/0193841X06291524.

Leyva, D., Barata, M., Snow, C., Weiland, C., Yoshikawa, H., Trevino, E., & Rolla, A. (2015). Teacher-child interactions in Chile and their associations with prekindergarten outcomes. *Child Development*, *86*, 661–694. http://dx.doi.org/10.1111/cdev.12342.

LoCasale-Crouch, J., Konold, T., Pianta, R., Howes, C., Burchinal, M., Bryant, D., ... Barbarin, O. (2007). Observed classroom quality profiles in state-funded pre-kindergarten programs and associations with teacher, program, and classroom characteristics. *Early Childhood Research Quarterly*, *22*, 3–17. http://dx.doi.org/10.1016/j.ecresq.2006.05.001.

Love, J. M., Harrison, L., Sagi-Schwartz, A., van IJzendoorn, M. H., Ross, C., Ungerer, J. A., ... Chazan-Cohen, R. (2003). Child care quality matters: How conclusions may vary with context. *Child Development*, *74*, 1021–1033 (doi: 0009-3920/2003/7404-0004).

Martinez-Beck, I. (2011). Introduction: Why strengthening the measurement of quality in early childhood settings has taken on new importance. In M. Zaslow (Ed.), *Quality measurement in early childhood settings* (pp. xviii–xxxiv). Baltimore, MD: Paul H. Brookes.

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., ... Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development*, *79*, 732–749. http://dx.doi.org/10.1111/j.1467-8624.2008.01154.x.

Maydeu-Olivares, A., Cai, L., & Hernandez, A. (2011). Comparing the fit of item response theory and factor analysis models. *Structural Equation Modeling*, *18*, 333–356. http://dx.doi.org/10.1080/10705511.2011.581993.

Mayer, R. E. (2002). Rote versus meaningful learning. *Theory Into Practice*, *41*, 226–233.

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Netherlands Consortium Kinderopvang Onderzoek [NCKO] (2011). *Pedagogische kwaliteit van de kinderopvang en de ontwikkeling van jonge kinderen: een longitudinale studie [Pedagogical quality of Dutch child care en de development of young children: a longitudinal study].* Amsterdam/Nijmegen: NCKO.

Pakarinen, E., Lerkkanen, M. -K., Poikkeus, A. -M., Kiuru, N., Siekkinen, M., Rasku-Puttonen, H., & Nurmi, J. -E. (2010). A validation of the classroom assessment scoring system in Finnish kindergartens. *Early Education & Development*, *21*, 95–124. http://dx.doi.org/10.1080/10409280902858764.

Perlman, M., Zellman, G. L., & Le, V. -N. (2004). Examining the psychometric properties of the early childhood environmental rating scale-revised (ECERS-R). *Early Childhood Research Quarterly*, *19*, 398–412. http://dx.doi.org/10.1016/j.ecresq.07.006.

Pianta, R., Howes, C., Burchinal, M., Bryant, D., Clifford, R., Early, D., ... Barbarin, O. (2005). Features of pre-kindergarten programs, classrooms, and teachers: Do they predict observed classroom quality and child-teacher interactions? *Applied Developmental Science*, *9*, 144–159. http://dx.doi.org/10.1207/s1532480xads0903_2.

Pre-COOL Consortium (2012). *Pre-COOL Cohortonderzoek. Technisch rapport tweejarigen onderzoek, eerste meting 2010–2011 [Pre-COOL cohort study. Technical report two-year-olds' cohort, first measurement wave 2010–2011].* Amsterdam: Kohnstamm Instituut.

Raver, C. C. (2004). Placing emotional self-regulation in sociocultural and socioeconomic contexts. *Child Development*, *75*(2), 346–353.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*, 68–78.

Singer, E., Nederend, M., Penninx, L., Tajik, M., & Boom, J. (2014). The teacher's role in supporting young children's level of play engagement. *Early Child Development and Care*, *184*, 1233–1249. http://dx.doi.org/10.1080/03004430.2013.826530.

Skibbe, L., Behnke, M., & Justice, L. M. (2004). Parental scaffolding of children's phonological awareness skills: Interactions between mothers and their preschoolers with language difficulties. *Communication Disorders Quarterly*, *25*(4), 189–203.

Slot, P. L., Leseman, P. P. M., Mulder, H., & Verhagen, J. (2013). *Handleiding CLASS Toddler [Manual CLASS Toddler].* Utrecht: Universiteit Utrecht, Onderwijsadvies en Training.

Slot, P. L., Leseman, P. P. M., Verhagen, J., & Mulder, H. (2015). Associations between structural quality aspects and process quality in Dutch early childhood education and care settings. *Early Childhood Research Quarterly*, *33*, 64–76. http://dx.doi.org/10.1016/j.ecresq.2015.06.001.

Sroufe, L. A. (2000). Early relationships and the development of children. *Infant Mental Health Journal*, *21*(1–2), 67–74.

von Suchodoletz, A., Fäsche, A., Gunzenhauser, C., & Hamre, B. K. (2014). A typical morning in preschool: Observations of teacher-child interactions in German preschools. *Early Childhood Research Quarterly*, *29*, 509–519. http://dx.doi.org/10.1016/j.ecresq.2014.05.010.

Taylor, B. M., Pearson, P. D., Peterson, D. S., & Rodriguez, M. C. (2003). Reading growth in high-poverty classrooms: The influence of teacher practices that encourage cognitive engagement in literacy learning. *The Elementary School Journal*, *104*, 3–28.

Thomason, A. C., & La Paro, K. M. (2009). Measuring the quality of teacher-child interactions in toddler child care. *Early Education & Development*, *20*, 285–304. http://dx.doi.org/10.1080/10409280902773351.

Tout, K., Starr, R., Soli, M., Moodie, S., Kirby, G., & Boller, K. (2010). The child care quality rating system (QRS) assessment: Compendium of quality rating systems and evaluations. *Child Trends & Mathematica Policy*http://www.acf.hhs.gov/programs/opre/cc/childcarequality

van Tuijl, C., & Leseman, P. P. M. (2007). Increases in the verbal and fluid cognitive abilities of disadvantaged children attending preschool in The Netherlands. *Early Childhood Research Quarterly*, *22*, 188–203. http://dx.doi.org/10.1016/j.ecresq.2007.02.002.

Vermeer, H. J., van IJzendoorn, M. H., de Kruif, R. E. L., Fukkink, R. G., Tavecchio, L. W. C., Riksen-Walraven, J. M., & van Zeijl, J. (2008). Child Care in the Netherlands: Trends in quality over the years 1995–2005. *The Journal of Genetic Psychology*, *169*, 360–385. http://dx.doi.org/10.3200/GNTP.169.4.360-385.

Vogel, C. A., Caronongan, P., Thomas, J., Bandel, E., Xue, Y., Aikens, N., ... Murphy, L. (2015a). *Toddlers in early head start: A portrait of 2-year-olds, their families, and the programs serving them (OPRE Report No. 2015–10).* Washington, DC: US Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation.

Vogel, C. A., Caronongan, P., Xue, Y., Thomas, J., Bandel, E., Aikens, N., ... Murphy, L. (2015b). *Toddlers in early head start: A portrait of 3-year-olds, their families, and the programs serving them (OPRE Report No. 2015–28).* Washington, DC: US Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation.

Zaslow, M., Anderson, R., Redd, Z., Wessel, J., Tarullo, L., & Burchinal, M. (2010). Quality, dosage, thresholds, and features in early childhood settings: A review of the literature (OPRE 2011-5). Retrieved from http://www.acf.hhs.gov/sites/default/files/opre/quality_review_0.pdf