Early Childhood Research Quarterly 54 (2021) 1–12

Contents lists available at ScienceDirect

Early Childhood Research Quarterly

Null relations between CLASS scores and gains in children's language, math, and executive function skills: A replication and extension study☆,☆☆

Paola Guerrero-Rosada[a,*], Christina Weiland[a], Meghan McCormick[b], JoAnn Hsueh[b], Jason Sachs[c], Catherine Snow[d], Michelle Maier[b]

[a] University of Michigan, United States
[b] MDRC, United States
[c] Boston Public Schools, United States
[d] Harvard Graduate School of Education, United States

ARTICLE INFO

Article history:
Received 22 July 2019
Received in revised form 22 July 2020
Accepted 23 July 2020

Keywords:
CLASS
Replication
Preschool quality
Baseline skills

ABSTRACT

General measures of process quality are widely used in the early childhood education (ECE) field. However, the evidence regarding associations between the most widely used process quality measure, the Classroom Assessment Scoring System (CLASS; Pianta et al., 2008), and children's school readiness gains during the preschool year is mixed. Using data collected during the 2016–2017 school year, we replicate prior work from the 2009–2010 school year which analyzed associations between CLASS scores and children's gains in language and executive function during the year when children were enrolled in a high-quality public prekindergarten program (Weiland et al., 2013). Additionally, we extend prior work by examining gains in numeracy skills and heterogeneous associations by children's skills at preschool entry. Participants were teachers in 42 preschool classrooms and a random sample of 307 children. Across linear, quadratic, and spline models, we found that none of the CLASS domains were associated with children's gains in vocabulary and executive function skills. We found no evidence of moderation by child baseline skills. We discuss future directions for measuring and analyzing process quality in prekindergarten.

1. Introduction

Approximately 69% of four-year-old children in the United States attend preschool and 44% are enrolled in publicly funded programs (Friedman-Krauss et al., 2018; Whitehurst & Klein, 2015). There is a strong consensus that children who attend preschool are better prepared for kindergarten compared to their peers who do not (Phillips et al., 2017). However, less is known about which specific features or active ingredients drive children's gains in preschool. Structural quality elements like teacher–child ratio and meeting safety standards appear to be necessary but not sufficient for high-quality preschool classroom experiences (Yoshikawa et al., 2013). Accordingly, in the search for the active ingredients in preschool, attention has shifted to process quality, defined in one very influential theory as the extent to which there are sensitive and responsive interactions between the teacher and children within the classroom (Burchinal et al., 2008; Mashburn et al., 2008; Pianta & Hamre, 2009). The leading measure of process quality is the Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2008). In recent years, this measure has been central to large-scale policy efforts to measure and improve programs nationwide, particularly in Head Start (U.S. Department of Health & Human Services, 2018). However, associations between the CLASS and

☆ Note: The research reported here was conducted as a part of a study funded by Arnold Ventures and by Grant <GN1>R305N160018<GN1> – 17 from the Institute of Education Sciences to MDRC with subcontracts to the University of Michigan, the Boston Public Schools, and the Harvard Graduate School of Education. The opinions expressed are those of the authors and do not represent views of the Institute, the U.S. Department of Education, or Arnold Ventures.

☆☆ Thanks to the Boston Public Schools, Brian Gold, Blaire Horner, Anne Taylor, the BPS Department of Early Childhood coaches and staff, the BPS Department of Research, the MDRC research team (Rama Hagos, Marissa Strassberger, Mirjana Pralica, Kelly Terlizzi, and Desiree Alderson), the Harvard Graduate School of Education research team (Sibyl Holland, Maia Gokhale, and the team of field-based data collection staff), the University of Michigan research team (Amanda Ketner, Lillie Moffett and Kehui Zhang), and grant project officer Caroline Ebanks.

* Corresponding author at: University of Michigan, 610 E. University Ave Room 1413, Ann Arbor, MI 48109, United States.
E-mail address: pguerre@umich.edu (P. Guerrero-Rosada).

https://doi.org/10.1016/j.ecresq.2020.07.009
0885-2006/

children's gains in preschool are modest and inconsistent across settings and studies (Burchinal, 2018).

In this article, we replicate and extend a prior study conducted in the Boston Public Schools (BPS) that examined whether higher process quality in preschool, as measured by the CLASS, was associated with gains in children's executive function skills and vocabulary across the preschool year (Weiland, Ulvestad, Sachs, & Yoshikawa, 2013). The prior study largely found null relations between CLASS scores and these outcomes. Given continued national interest in the CLASS as both an indicator of preschool program quality and a tool for program improvement, we extend the prior work and add to the broader literature in four ways. First, we use data collected during the 2016–2017 school year in the same original study district (the prior study used data from the 2009–2010 school year). Using the same analytical approach across years allows for minimal differences in results due to model specification, since we test the same number of parameters within the same district context (IES & NSF, 2018; Schauer, 2018). Moreover, by examining the same district context, we account for differences in at least two frequent confounders associated with process quality, namely, curriculum implementation and structural quality. Second, we explore whether high-quality interactions predict gains in children's math skills at the end of the preschool year. Math was not included in the prior study due to data constraints and may be particularly sensitive to higher instructional quality during preschool (Clements & Sarama, 2011; Nguyen et al., 2016). Third, we explore whether children's baseline skills moderate associations between classroom quality and children's gains in numeracy, vocabulary and executive function skills. Some children gain more than others from preschool in general (Phillips et al., 2017) and understanding if those with lower baseline skills gain more from higher process quality adds to the field's understanding of heterogeneity in benefits of early childhood programs (Bloom & Weiland, 2015).

Finally, our goals of replication and extension help meet calls across multiple fields for use of robustness checks in empirical work and for conducting replication studies to determine whether key findings hold when examined across different data sets and demographic subgroups (Duncan, Engel, Claessens, & Dowsett, 2014; Ioannidis, 2005; Makel & Plucker, 2014; Pashler & Harris, 2012). There has been considerable gentrification in Boston between the time of the original study and the current one, a demographic trend that has substantially increased the proportion of higher-income families enrolling their children in BPS (Lima & Melnik, 2014). Identifying whether the largely null results from the prior study on a 2009–2010 sample hold in our current 2016–2017 sample helps to identify the degree to which the relations we detected are sample invariant or not. Ultimately, this evidence can be used to make better decisions about how the CLASS should be used to support program improvement in the Boston context and in other prekindergarten contexts.

1.1. Associations between high quality and children's developmental gains in preschool

The early childhood education (ECE) literature has used the Theory of Effective Interactions to explain general indicators of classroom processes that measure teachers' responsivity and sensitivity as mechanisms for promoting learning and development (Pianta & Hamre, 2009). One example used extensively in the ECE literature is the Classroom Assessment Scoring System (CLASS; Pianta et al., 2008) and its three domains: Instructional Support, Classroom Organization, and Emotional Support. Instructional Support measures the extent to which teachers provide opportunities and feedback to enhance higher-order thinking skills and language. Emotional Support measures the extent to which the classroom climate is positive for children's social interactions and teachers are responsive to children's needs. Finally, Classroom Organization measures the teachers' management of time, behavior, and attention in the classroom (Pianta & Hamre, 2009; Pianta et al., 2008).

1.2. The CLASS theoretical framework

The CLASS theoretical framework posits that daily back-and-forth exchanges among children and teachers are the primary mechanism driving children's learning. These interactions occur throughout the day and create opportunities for children to engage in instructional and social challenges, thus supporting growth in their cognitive and emotional skills (Hamre, Hatfield, Pianta, & Jamil, 2014; Mashburn et al., 2008; Pianta & Hamre, 2009). Socio-emotional features of interactions, such as teachers' sensitivity or regard for students' perspectives, are expected to promote children's higher-order thinking skills. Instructional aspects of interactions, such as the quality of feedback, or language modeling, are expected to promote children's early literacy, language, and other school readiness skills (Early, Maxwell, Ponder, & Pan, 2017; Hamre et al., 2012; Mashburn, Downer, Hamre, Justice, & Pianta, 2010).

1.3. The CLASS empirical background

Interventions based on the CLASS theoretical model and classroom quality measure have been shown to improve teacher practice (Early et al., 2017; Pianta et al., 2017), yet associations between the CLASS domains and children's academic gains in preschool are usually modest or null (Burchinal, 2018). For the outcomes examined in the current study – specifically, math, language, and executive function – the evidence is mixed. Some studies have found Instructional Support to be a statistically significant predictor of gains in language, literacy, math, and inhibitory control skills (Hamre et al., 2014; Mashburn et al., 2008). However, a recent meta-analysis including 35 peer-reviewed studies found no associations between Instructional Support and children's gains in any of these outcomes (Perlman et al., 2016). Classroom Organization was the only CLASS domain examined in this meta-analysis that consistently predicted inhibitory control gains across studies, with a small pooled correlation coefficient of 0.06 ($p < 0.05$; Perlman et al., 2016).

Noting these findings, some researchers have examined whether the association between process quality and children's gains in preschool is non-linear (Weiland et al., 2013). To explore this alternative, researchers have used quadratic models and tested whether there are thresholds of process quality that must be observed in order to see corresponding gains in students' skills. This modeling approach has yielded some moderate-sized associations between the CLASS domains and gains in student outcomes at higher levels of quality (Burchinal, Vandergrift, Pianta, & Mashburn, 2010). Even so, other studies using this approach have also found similar null or small associations (Burchinal et al., 2008; Zaslow et al., 2011).

For example, in some studies using non-linear approaches (i.e., threshold analysis or spline models), Instructional Support has been a statistically significant predictor of children's gains in math, vocabulary (Burchinal et al., 2010), and inhibitory control (Weiland et al., 2013) at both low-moderate and high levels of quality. In others, relations between Instructional Support and children's gains in a variety of other language outcomes and inhibitory control have been null (Burchinal, Vernon-Feagans, Vitiello, & Greenberg, 2014; Hatfield, Burchinal, Pianta, & Sideris, 2016). Across studies, standardized associations have ranged from −0.20 *SD* (inhibitory control) to 0.08 *SD* (expressive language) in low-quality settings and from 0.19 *SD* (inhibitory control) to 0.34 *SD* (math) in high-quality settings. Overall in the literature, children in low-quality

classrooms gain less when compared to similar children in high-quality settings (Burchinal et al., 2010; Weiland et al., 2013).

Regarding Classroom Organization and Emotional Support, the evidence is also inconsistent. Burchinal et al. (2010) found that Emotional Support was not a predictor of math or language gains in low- or high-quality settings. However, for children attending high-quality preschool, Hatfield et al. (2016) found positive and statistically significant associations with gains in phonological awareness (Emotional Support $d = 0.49$; Classroom Organization $d = 0.84$) and inhibitory control (Emotional Support $d = 0.43$). In contrast, Burchinal et al. (2014) found that children attending high-quality settings gained less in their math skills than their peers attending low-quality settings.

1.4. Explanations for the patterns of mixed results

There are many potential explanations for these patterns of mixed findings. One that is particularly relevant to our study is that the curriculum used and the content children are exposed to in the preschool classroom may be more influential in supporting gains in children's math, language, and executive function skills than the more general instructional practices measured by the CLASS. For example, teachers in the current study sample implemented the mathematics curriculum, Building Blocks, which has a strong track record in improving children's math skills (Clements & Sarama, 2008, 2011). Moderate fidelity to this curriculum across the school year might have provided children with enough early mathematics practice and instruction to promote their learning, even if a teacher is low to middling in instructional quality on the days they were observed on the CLASS. Note that a strength of our current replication and extension study is that we do not expect differences in curriculum to confound linkages between the CLASS and gains in student outcomes in the original study versus the current study, since the same literacy and math curricula have been in place in the BPS prekindergarten program since the 2007–2008 school year.

Another possible explanation for these mixed findings linking CLASS domains and student outcomes is selection bias. More advantaged families generally select into higher-quality settings and less advantaged families, into lower-quality classrooms (Chaudry, Morrissey, Weiland, & Yoshikawa, 2017; Hillemeier, Morgan, Farkas, & Maczuga, 2013; Perlman et al., 2016). The previous literature has attempted to adjust for this selection issue using demographic control variables, but there may be unobservable family characteristics that affect teacher-child interactions and children's developmental gains simultaneously. For instance, there is evidence that children's math and reading baseline skills are predictive of teachers' CLASS scores at the end of the year (Pakarinen, Lerkkanen, Poikkeus, Siekkinen, & Nurmi, 2011). If so, the variability in children's skills within and between classrooms could limit the extent to which significant associations between gains in children's outcomes and CLASS scores will generalize across different samples. Although the non-random nature of children's selection into classrooms is also a feature of our current study, we do examine sorting by child demographic characteristics and skills into lower and higher quality classrooms to try to understand the extent to which selection bias poses a problem for this analysis. As a preview, we find little to no evidence of sorting by children's baseline skills or demographic characteristics (see Section 3 for more details and findings).

Third, process quality may be more important for students who start the year with lower skills relative to their peers. One study (Keys et al., 2013) used secondary data from the Study of Early Child Care and Youth Development (SECCYD-NICHD), the Early Childhood Longitudinal Study – Birth Cohort (ECLS-B), the National Center for Early Development and Learning's Multi-State Study of Pre-Kindergarten (NCDEL-NIEER) and the Early Head Start Study (EHS) to test whether cognitive and attention skills, as well as problem behaviors, moderated the main effects of Instructional Support in linear models. Results were null. But supporting this somewhat under-explored possibility, studies conducted in Head Start settings have shown that children with lower baseline skills make greater gains in preschool, in general, than their peers (Bitler, Hoynes, & Domina, 2014; Bloom & Weiland, 2015). Along similar lines, another study found that Head Start boosted gains in math only for children receiving low parental pre-academic stimulation in language and literacy (Miller, Farkas, & Duncan, 2016). It is possible that process quality might operate similarly and matter more for children who start the preschool year already disadvantaged due to their skill entry levels – a possibility we examine in our study across the full set of CLASS domains.

1.5. Present study

In the present study, we replicate and extend a study conducted during the 2009–2010 school year that included 414 children attending the BPS prekindergarten program (Weiland et al., 2013). In that analysis, Weiland and colleagues examined linear and non-linear associations between structural and process quality and gains in children's vocabulary and inhibitory control during the preschool year. The authors found non-linear associations between the three CLASS domains and inhibitory control measured with the Pencil Tapping task (Diamond & Taylor, 1996), showing that associations between each CLASS domain and inhibitory control were stronger at higher levels of classroom quality. In our replication, we examine process quality, as measured by the CLASS, only due to data limitations and to limited variation in the program's structural characteristics related to standardized teacher educational requirements, class size, etc. We follow the original study's analytical approach as closely as possible, with the goal of minimizing differences in results due to modeling decisions. In doing so, we add to the broader literature by examining three research questions. The first question serves to replicate prior work and the second and third questions extend the prior research:

1 Is higher classroom process quality as measured by the CLASS associated with gains in children's executive function skills and vocabulary during the 2016–2017 prekindergarten year?
2 Is higher classroom process quality as measured by the CLASS associated with gains in children's math skills during the 2016–2017 prekindergarten year?
3 Are the associations between classroom process quality as measured by the CLASS and children's gains in numeracy, vocabulary, and executive function skills moderated by these skills at preschool entry?

2. Method

2.1. Participants and setting

We used data from the 2016 to 2017 year of the ExCEL P-3 Study (Expanding Children's Early Learning: Sustaining Gains from Preschool to Third Grade) which is part of the broader Institute of Education Sciences Early Learning Network. Our sample included 307 children (50% female) nested in 42 classrooms in 20 public schools offering a prekindergarten program. Sample members were diverse in their home language, income, ethnicity, and race as well as executive function, vocabulary, and math skills at the start of the preschool year (see Table 1).

We drew our sample from the broader population of BPS elementary schools that had prekindergarten classrooms. Specifically, we randomly selected 25 schools from the full set of 76 schools that

Table 1
Descriptive statistics: children's demographic characteristics, fall skills, and spring skills.

		N	Mean or %	SD	% Missing
Demographic characteristics					
	Female		50.16	–	0
	Eligible for Free or Reduced Lunch (FRPL)		57.98	–	0
	Dual Language Learner (DLL)		53.75	–	0
	Asian Pacific Islander		15.64	–	0
	Black		19.87	–	0
	Latinx		30.29	–	0
	Other		6.84	–	0
	White		27.36	–	0
Fall measures					
	PPVT – Raw Score	297	73.50	28.44	3.25
	WJ Applied Problems Raw Score	298	12.52	5.18	2.93
	Forward Digit Span	298	3.13	1.06	2.93
	Hearts and Flowers	251	0.59	0.18	18.24
Spring measures					
	PPVT – Raw Score	292	87.02	27.22	4.89
	WJ Applied Problems Raw Score	291	15.69	4.73	5.21
	REMA Raw Score	292	20.28	9.49	4.89
	REMA T Score	292	37.22	6.38	4.89
	REMA IRT	292	−2.58	1.29	4.89
	Forward Digit Span	292	3.49	1.02	4.89
	Hearts and Flowers	261	0.68	0.21	14.98

Note: $N = 307$ for children's demographic characteristics.

offered the public prekindergarten program. Four of the selected schools declined participation and one was designated as a pilot school for developing new measures and thus was not included in the present study's sample, leaving us with a sample of 20 participating schools. On average, across the current sample schools, 48% of students were eligible for free or reduced-price lunch, 49% of students were Dual Language Learners (DLL), 26% were Black, 16% were White, 46% were Hispanic, 9% were Asian, and 3% were mixed race or another race. About 40% of third-grade students in study schools met or exceeded expectations on the 2015–2016 state English/Language Arts exam, and 45% met or exceeded expectations on the state math exam. Study schools are generally representative of the broader population of BPS schools offering a prekindergarten program, but had lower proportions of Black students (32% at the district level) and higher proportions of students meeting or exceeding expectations on the 2015–2016 ELA exam (36% at the district level).

Compared with the sample in the original study (see Table 1), fewer children in the 2016–2017 sample were eligible for FRPL (68% in 2009 versus 58% in 2016), identified as Hispanic (43% in 2009 versus 30% in 2016), or identified as Black (28% in 2009 versus 20% in 2016). More participants in the current sample were DLLs (48% in 2009 versus 54% in 2016), identified as Asian (11% in 2009 versus 16% in 2016), or identified as mixed race (3% in 2009 versus 7% in 2016). Another sample difference is that 13% of children in the 2009 sample were classified as having a special need. Due to a funding requirement for the current study (Hsueh, 2016) we did not recruit children with special needs in the 2016–2017 sample.

All children in the current sample were enrolled in the BPS prekindergarten program. The BPS prekindergarten program is free to families who live in the city of Boston, is full-day (6.5 h), and is open to age-eligible children (i.e., to children who turn 4 by the September 1 cutoff date for a given school year). Almost all the teachers in our sample (92%) reported using the *Focus on K1* curriculum. *Focus on K1* is based on the language and literacy-focused curriculum, Opening the World of Learning (Schickedanz & Dickinson, 2004) and the mathematics curriculum Building Blocks (Clements & Sarama, 2008; Sarama & Clements, 2004). The district also designed additional curriculum components – Storytelling/Storyacting (oral language-focused) and Thinking and Feedback (critical thinking and oral language focused) – meant to be implemented from prekindergarten through second grade to better align instruction across grades. To support implementation, teachers had opportunities to participate in training on the curriculum and on-going coaching with district staff.

All sample teachers held a bachelor's degree and 90% held a master's degree. Overall, 62.5% of those degrees were specific to ECE. On average, sample teachers had nine years of preschool teaching experience ($SD = 7.36$ years). All prekindergarten teachers were required to attain an early childhood license from the Massachusetts Department of Elementary and Secondary Education and were paid on the same scale as K-12 teachers.

2.2. Procedures

All the Institutional Review Boards in the organizations conducting the ExCEL study approved the human subjects plan before our team began research activities. Following children's enrollment in BPS prekindergarten in the Fall of 2016, we collected active consent for the prekindergarten students enrolled in the participating classrooms. Overall, 81% of families consented for their child to participate. Within each classroom with a maximum of 22 students (adult–child ratio was 2:22), we randomly selected approximately half of consented children to participate in the data collection, for a total sample size of 307 (range of 4 to 10 per classroom, average = 7.14). Nine children did not participate in assessments during the Fall and 15 children were not tested during the Spring.

Between September and November of 2016, we assessed children's baseline executive function, vocabulary, and math skills. Children were re-assessed in the Spring of 2017. Before each data collection period, child assessors were trained over the course of five days by a master trainer with experience in field-based studies. Following the training, each assessor demonstrated reliability first by engaging in a mock assessment with an adult acting as a child, and second by participating in a practice assessment with a child in the field who was not enrolled in the study. A field supervisor further observed 10% of assessments during the data collection process to ensure continued high-quality administration.

At the beginning of each assessment session, the assessor used the Pre-language Assessment Scale (preLAS) Simon Says and Art Show tests (Duncan & De Avila, 1998) to determine the testing administration language and as a warm-up to the assessment

battery. The preLAS is a measure of pre-literacy skills and an individual's proficiency in English. The preLAS has been used as a screener to identify whether a child completes subsequent assessments in English or Spanish, based on the number of items answered correctly (Barrueco, López, Ong, & Lozano, 2012). If the child answered 5 or more items on the preLAS incorrectly and the parent indicated that Spanish was their home language (33 participants), the rest of the testing battery except the Peabody Picture Vocabulary Test-IV (PPVT-IV) was administered in Spanish. Among the 307 children in the current study sample, 20 (6.5%) completed the assessments in Spanish in the fall and 6 completed the assessment in Spanish (2%) in the spring.

During the Winter of 2017, each classroom was videotaped for two hours during two visits. Visits were scheduled in advance with teachers. In addition to the lead teacher, a paraprofessional was present in 88% of the classrooms on visit days. Observers participated in a two-day training to learn the CLASS measure and then established reliability on a set of master codes created by the test developers. Coders started coding the tapes once instructional time began. As recommended by the measure's protocol (Pianta et al., 2008), coders used cycles of 20 min for observing and 10 min for scoring, which they repeated 4 times for each observation. Scores across the four segments were first averaged to calculate observation-specific scores and then the scores across observations were averaged to generate one overall score for each classroom. The CLASS was live-coded in the original study. Recent evidence has shown that although video-coded classrooms obtain slightly lower scores than live-coded classrooms, the predictive validity of the tool does not vary across the methodologies (Curby, Johnson, Mashburn, & Carlis, 2016). The team double-coded 20% of the observations to assess interrater reliability. The final ICCs representing interrater reliability were 96% for Emotional Support, 94% for Classroom Organization, and 88% for Instructional Support. We also did a drift check wherein coders had to code a master tape every three weeks to ensure they were still reliable before continuing to code tapes.

2.3. Measures

2.3.1. Classroom process quality

General classroom process quality was measured using the Classroom Assessment Scoring System (CLASS) PreK (Pianta et al., 2008). This observational tool measures three domains of teacher-child interactions: Emotional Support, Classroom Organization, and Instructional Support. Emotional Support is a composite measure of four subscales – positive and negative climate, sensitivity and regard for students' perspectives. Classroom Organization includes measures of behavior management, productivity, and instructional learning formats. Instructional Support includes concept development, language modeling, and quality of feedback. All the dimensions are directly scored on a 7-point scale, except for negative climate which is reverse-coded. The CLASS and these three constructs show good psychometric validity in the literature and prior studies examining associations between quality and children's outcomes at different levels of quality have used this same three-factor structure (Burchinal et al., 2014; Hatfield et al., 2016; Leyva et al., 2015; Weiland et al., 2013). In our study, we empirically assessed the psychometric properties of the CLASS three-factor model using confirmatory factor analysis. Consistent with prior literature, our model demonstrated a good fit to the data ($x^2_{(29)} = 32.62$, $p = 0.29$, CFI = 0.99, TLI = 0.98, RMSEA = 0.06, SRMR = 0.05) and the three subscales also had good internal consistency (Emotional Support, $\alpha = 0.87$; Classroom Organization, $\alpha = 0.88$; Instructional Support, $\alpha = 0.92$). These results were consistent with the original study, in which a three-factor solution had adequate fit to the data ($x^2 = 57.73$, $p < 0.01$, CFI = 0.97, TLI = 0.94, RMSEA = 0.11, SRMR = 0.06). In our linear and quadratic models, we used continuous measures of Emotional Support, Classroom Organization and Instructional Support. Cut points for our spline models are explained in the analytic section.

2.3.2. Receptive vocabulary

The PPVT-IV has been normed and used widely in diverse samples of children in the U.S (Puma, Bell, Cook, Heid, & U.S Department of Health and Human Services, 2010), and it has shown qualitative and quantitative validity properties (Dunn & Dunn, 2007). The test–retest reliability ranges from 0.92 to 0.96. The PPVT IV measures children's vocabulary acquisition in standard American English. It requires children to choose (verbally or nonverbally) which of four pictures best represents a stimulus word. In our primary analysis, we used the raw score total as our outcome measure. However, we report models using the age-standardized score versions, interpreted relative to national norms where a score of 100 represents the national average in the Appendix (see Table A2). As explained in the previous section, we assessed all children on the PPVT in English regardless of their results on the PreLAS language screener in order to obtain an English receptive language score for the full sample.

2.3.3. Math skills

To assess children's early math skills, we used the Woodcock–Johnson Applied Problems III (Woodcock, Mather, McGrew, & Wendling, 2001) subtest and the Research-based Early Math Assessment (REMA; Clements, Sarama, & Liu, 2008; Weiland et al., 2012).

The Woodcock–Johnson Applied Problems subtest requires children to perform relatively simple calculations to analyze and solve arithmetic problems. Its estimated test–retest reliability for 2- to 7-year-old children is 0.90 (Woodcock et al., 2001) and it has been nationally normed and used with diverse populations of children (Gormley, Gayer, Phillips, & Dawson, 2005; Wong, Cook, Barnett, & Jung, 2008). The team assessed Spanish-speaking children who did not pass the PreLAS language screener using the equivalent Spanish language version of the assessment from the Batería III Woodcock Muñoz (Woodcock, Munoz-Sandoval, Ruef, & Alvarado, 2005). We present results using the raw score of the measure, but models using the age-standardized version of the Applied Problems scores are reported in the Appendix (see Table A2). In our sample, the majority of children completed the test in English (6.5% of the sample completed the assessment in Spanish in the fall and 2% completed it in Spanish in the spring). The Woodcock–Muñoz assessment in Spanish follows similar norming strategies to the Woodcock–Johnson English version and allows for combining scores across both English and Spanish in the same sample.

We also used the Research-based Early Mathematics Assessment to assess math skills (REMA; Clements & Sarama, 2011). The REMA is a hands-on, one-on-one assessment of children's early math skills (e.g., numeracy, geometry, operations, spatial reasoning). The alpha reliabilities of the test subscales range from $r = 0.89$ (number) to 0.71 (geometry). We present results using the REMA raw score, as well as results from models using the t and IRT scores in the Appendix (see Table A2). The IRT score takes item difficulty into account, while the t score also reflects the level of difficulty of the strategy a child used to answer a given question. We did not assess children on the REMA during the Fall and thus use the Woodcock–Johnson Applied Problems as a baseline covariate for all models examining math skills as an outcome.

2.3.4. Executive function

We used two different measures to capture components of children's executive function. The first – the Forward Digit Span

Assessment – measures children's working memory. It requires that children repeat several series of numbers in rapid succession, with an increasing number of digits presented once the child has successfully repeated a prior sequence (Wechsler, 1974). This test is widely used and nationally normed. We used the categorical score for Forward Digit Span (FDS), which represents the sequence with the highest number of digits that the child repeated accurately. Recent clinical evidence supports the discriminative properties of this task in differentiating typically developing children from those with specific working memory impairments (Giofrè, Stoppa, Ferioli, Pezzuti, & Cornoldi, 2016). Additionally, FDS has high correlations with Backward Digit Span and other EF tasks, and has shown good test–retest reliability in samples of prekindergarten children ($r = 0.80$; Muller, Kerns, & Konkin, 2012).

We also used the Hearts and Flowers task, a measure of inhibitory control and working memory (Davidson, Amso, Anderson, & Diamond, 2006), to assess executive function. This test combines the cognitive demand of the Simon Says (Duncan & De Avila, 1998) and spatial Stroop tasks (Hilbert, Nakagawa, Bindl, & Bühner, 2014). During congruent trials, children need to obey the rule: "Press on the same side as the stimulus" and during incongruent trials, children follow the opposite rule: "Press on the side opposite the stimulus" (Wright & Diamond, 2014). We used the Incongruent Trial score because it best approximates the Pencil Tap measure (Diamond, Barnett, Thomas, & Munro, 2007) used in the original study (Weiland et al., 2013) that we are replicating and extending. Prior studies have shown that the Hearts & Flowers task predicts behavioral and academic outcomes, thus supporting its criterion validity (Camerota, Willoughby, & Blair, 2019; Raver et al., 2011). The main difference between the Pencil Tap and the Hearts and Flowers tasks is that the Pencil Tap task requires children to inhibit a gross motor behavior in response to an auditive stimulus whereas the Hearts and Flowers incongruent trial requires children to inhibit the tactile selection of visual stimuli in a digital screen. Both assessments aim to capture children's inhibitory control, with the Hearts and Flowers assessment automating much of the testing process and thus reducing the probability of assessor error.

2.3.5. Child-level covariates

We accessed administrative data from the school district to determine children's race/ethnicity, eligibility for free or reduced-price lunch (FRPL), gender, and DLL status. We defined a set of indicators to describe children's race/ethnicity (Black, Hispanic, Asian, or Other Race/Ethnicity with White as the reference group). We used dichotomous indicators to capture whether each child was eligible for FRPL, was female, or was a DLL (determined based on parent's report that a language other than English was spoken at home, was the language most often spoken by the student, or was the student's first language). Finally, we used the child's birth date to calculate age at the time of the Fall 2016 assessment.

2.3.6. Classroom level covariates

Following the prior Boston study (Weiland et al., 2013), we used school administrative data to identify the number of children enrolled in each classroom at the beginning of the year and created a dichotomous indicator for classrooms with more than 20 enrolled students. We used teacher surveys conducted during Fall 2016 to create a dichotomous variable set equal to one if the teacher had a master's degree and 0 otherwise.

2.4. Analytical approach

2.4.1. Missing data

Overall, there was a relatively low amount of missing data across study variables. Two classrooms with 17 assessed children total were missing CLASS data because their teacher refused to participate in videotaping. For child outcomes, missingness was 2.9% during the Fall and 4.9% during the Spring for PPVT, Woodcock–Johnson Applied Problems, and Forward Digit Span assessments. There were no missing child-level covariates. Since we did not find evidence for systematic differences between the children missing and not missing data, we imputed 100 datasets using multivariate normal regression with Stata 15 (Graham, 2009, 2012; van Ginkel, Linting, Rippe, & van der Voort, 2020), with outcomes, covariates and the CLASS subscales included in the imputation model. Following procedures recommended by von Hippel (2009), we imputed quadratic CLASS variables, but we did not impute other interactions of interest (i.e., baseline skills and CLASS interactions as in RQ 3) to maintain a clear definition of our skills subgroups. To evaluate if our results were sensitive to our missing data patterns, we also present results using complete case analysis and results from an alternative multiple imputation model excluding outcomes as robustness checks.

2.4.2. RQ 1 and RQ 2: regression models with robust standard errors

To address our first two research questions which focus on the relations between process quality and gains in children's vocabulary, executive function, and math skills in preschool, we used multiple regression. Specifically, we fit the following model, separately for each outcome:

$$Outcome_{ij} = \beta_0 + \beta_1 Quality_j + \gamma_{ij} + \delta_j + \varepsilon_{ij}, \quad (1)$$

where i denotes child and j denotes classroom; β_1 is the coefficient of interest for the relevant classroom quality indicator (Emotional Support, Instructional Support, or Classroom Organization); γ_{ij} represents a vector of child-level demographic characteristics (including children's baseline measures for each corresponding outcome, race/ethnicity, eligibility for free or reduced lunch, gender, and DLL status); δ_j represents a vector of classroom indicators for class size (greater than 20) and whether the teacher has a master's degree; and ε_{ij} is a child-level error term. Following the precedent set by Weiland et al. (2013), we used robust standard errors to adjust for clustering at the classroom level in our models, since the nesting can introduce heteroskedasticity in model residuals and bias the standard errors (White, 1980).

Consistent with prior literature (2014, Burchinal et al., 2010; Hatfield et al., 2016; Weiland et al., 2013), we also modified Eq. (1) in two ways to examine non-linear associations between the key predictor and each outcome. First, we added a quadratic term for quality to Eq. (1). Then, we intended to fit the following spline regression models, using empirically defined and conceptually defined cut-points:

$$Outcome_{ij} = \beta_0 + \beta_1 LowQuality_J + \beta_2 HighQuality_j + \gamma_{ij} + \delta_j + \varepsilon_{ij}, \quad (2)$$

where β_1 and β_2 are the coefficients of interest, capturing low to moderate and high levels of quality respectively. All other terms are consistent with Eq. (1). Spline models are a linear regression approach in which the slope of a line is permitted to change at a given threshold but the model intercept is not (Marsh & Cormier, 2002). We use the simplest case of spline model, where the number of segments and the location of the knots (or cut points) are assumed to be known. To choose our cut points, we followed a theoretical criterium where scores above 5.0 are considered high-quality and scores below lower quality (Pianta et al., 2008). Given the range restrictions in Instructional Support, we adhered to 2.75 as a cut point for low and moderate quality, as used in prior literature (Mashburn, 2017; Zaslow et al., 2011). These theoretical cut points are the same as those used in the original Boston study (Weiland et al., 2013). Second, we intended to use basic calculus to estimate the inflection points in our quadratic models – i.e., the point at which the slope of the quadratic relationship between the

Table 2
Classroom characteristics.

	Mean or %	SD	Min	Max
CLASS – process quality				
Emotional Support	5.57	0.60	3.97	6.65
Instructional Support	3.22	0.63	2.25	4.50
Classroom Organization	5.49	0.58	3.67	6.62
Classroom structural characteristics				
Teacher has a master's degree	90%			
More than 20 students	12%			

Note: $N = 40$ classrooms.

CLASS measures and the child-level outcome changes in sign, as illustrated in the original study (Weiland et al., 2013). We did not use this approach as planned, since we did not find statistically significant quadratic associations.

For answering question 3 – whether associations between classroom quality and children's gains in numeracy, vocabulary, and executive function skills were moderated by the children's baseline skills – we created dichotomous indicators for whether the child was in the lowest quartile of the sample in their baseline measure of PPVT, Woodcock–Johnson Applied Problems, Forward Digit Span, or Hearts and Flowers, separately for each measure. We included the lowest quartile indicator for each outcome in Eq. (1), along with an interaction term between quality and the lowest quartile indicator. We did not explore moderation in our non-linear models due to our relatively small sample size (we return to this issue in the limitations section). To explore whether our moderation results are sensitive to modeling/coding decisions, we also estimated two alternative specifications for examining moderation: 1) using the lowest quartile indicator and excluding baseline scores; and 2) using the continuous baseline scores as a moderator only. We present these robustness checks in Appendix (see Tables A3 and A4).

3. Results

3.1. Descriptive statistics

As shown in Table 2, CLASS scores ranged from moderate to high in Emotional Support and Classroom Organization, and from low to moderate in Instructional Support. Overall, 88% of children attended classrooms with high Emotional Support ($N = 38$ classrooms, or 90% of classrooms), 82% of children were enrolled in classrooms with high Classroom Organization ($N = 36$ classrooms, or 86% of classrooms), and 61% attend classrooms with moderate Instructional Support ($N = 31$ classrooms, or 74% of classrooms). Compared with the original study, the percentage of classrooms with ratings falling below the quality cut points was similar for the three CLASS domains (Emotional Support, 14% in the original study and 10% in this replication; Classroom Organization, 13% in the original study and 14% in this replication; and Instructional Support, 28% in the original study and 26% in the replication).

We also conducted descriptive analysis aimed at identifying whether there was selection on demographic characteristics and/or baseline test scores into higher versus lower quality classrooms, based on our conceptually defined quality cut-offs (see Appendix, Table A1). For parsimony, we focused on Instructional Support and found little sorting by child demographic characteristics. We also found that there were no statistically significant differences in the baseline skills of students attending classrooms with low or moderate Instructional Support.

3.2. RQ1: relations between classroom process quality and gains in children's executive function and vocabulary skills

As shown in Table 3, none of the CLASS measures were statistically significantly associated with gains in children's working memory and vocabulary during the 2016–2017 prekindergarten year in either the linear or quadratic models. These results are consistent with Weiland et al.' findings (2013).

Given that none of the CLASS domains predicted gains in vocabulary, working memory, or inhibitory control in our quadratic models, we did not estimate spline models with empirically determined cut points as we originally intended. Using conceptually defined cut points in spline models, we found that our quality measures did not predict children's gains in these skills regardless of the classroom's quality level (see Table 4). These results differed from the original study, in which Emotional Support predicted fewer vocabulary gains in lower quality classrooms ($d = 0.30$; Weiland et al., 2013).

3.3. RQ2: relations between classroom quality and gains in children's math skills

As shown in Table 3, we found no linear or quadratic associations between any of the quality measures and children's gains in math for either of our two math outcomes. When using conceptually defined cut points to estimate spline regression models, we likewise found that the CLASS domains did not predict children's gains for either math measure in lower or higher quality classrooms (see Table 4).

3.4. RQ 3: moderation by child baseline skills

We extended our linear models by examining whether associations between the CLASS subscales and gains in language, math, and executive functioning skills varied by children's baseline skill levels, defined by entering prekindergarten in the lowest quartile for a given skill versus entering with a higher-level skill. We found no evidence of moderation (see Table 5).

3.5. Robustness checks

We conducted four sets of robustness checks to test whether our results were sensitive to our modeling or measurement choices. First, since we reported raw outcome scores in our main approach, we present results using standardized scores when these are available. Second, we examined whether our moderation results were sensitive to our modeling decisions via two different checks described below. Third, we fit our primary models using complete case analysis instead of multiple imputation (see Appendix, Tables A5–A7) and using an alternative multiple imputation model excluding outcomes (see Appendix, Tables A8–A10).

Regarding the first check, standardized scores were available for the PPVT, W-J Applied Problems, and the REMA (see Appendix, Table A2). We found that the magnitude, direction, and p-values were consistent across raw and standardized scores for these measures. For the second check, we first refit our moderation models excluding the binary indicator for lowest quartile and used the continuous baseline score as a moderator (see Appendix, Table A3). We then fit additional models including the lowest quartile indicator but excluding baseline scores (see Appendix, Table A4). Results were robust across these different modeling approaches. Finally, we found that our results were largely robust to how missing data issues were addressed (see Appendix, Tables A5–A10). There were a few exceptions in that some relations that were null in both of our missing imputation approaches were statistically significant in

Table 3
Results of regressing receptive vocabulary, executive function, and math measures on classroom quality indicators.

	PPVT – IV		Forward Digit Span		Hearts and Flowers		WJ – applied problems		REMA	
	Linear	Quadratic	Linear	Quadratic	Linear	Quadratic	Linear	Quadratic	Linear	Quadratic
Instructional Support	−2.44	22.53	−0.02	0.52	−0.04	−0.22	0.22	4.15	−0.25	8.20
	(1.75)	(20.04)	(0.09)	(0.96)	(0.02)	(0.21)	(0.33)	(3.46)	(0.62)	(6.51)
Instructional Support^2		−3.68		−0.08		0.03		−0.58		−1.25
		(3.01)		(0.14)		(0.03)		(0.51)		(0.96)
Classroom Organization	0.09	−11.64	0.04	0.68	−0.03	−0.25	0.07	1.10	0.36	14.08
	(2.00)	(23.07)	(0.11)	(1.13)	(0.02)	(0.24)	(0.36)	(4.16)	(0.68)	(7.31)
Classroom Organization^2		1.14		−0.06		0.02		−0.10		−1.33
		(2.18)		(0.11)		(0.02)		(0.41)		(0.72)
Emotional Support	1.22	−2.64	0.10	−0.19	−0.03	0.02	0.17	2.24	0.27	12.75
	(1.86)	(25.61)	(0.10)	(1.29)	(0.02)	(0.28)	(0.32)	(4.21)	(0.64)	(7.99)
Emotional Support^2		0.36		0.03		−0.01		−0.19		−1.15
		(2.30)		(0.12)		(0.03)		(0.39)		(0.75)

Note: All models control for children's age and baseline score on the requisite outcome, a set of binary indicators for child gender, free or reduced lunch status, home language, class size, and whether the teacher has a master's degree, and a categorical variable indicating the children's race/ethnicity. Robust standard errors (in parentheses) were used to adjust for clustering at the teacher level. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table 4
Results of spline regression models for receptive vocabulary, executive function and math measures using conceptually defined classroom quality levels.

		Low-quality	High-quality
Vocabulary – PPVT			
	Emotional Support	1.89	1.26
		(7.13)	(2.16)
	Classroom Organization	−0.52	0.74
		(5.27)	(2.67)
	Instructional Support	5.77	−3.31
		(9.54)	(2.24)
Inhibitory Control – FDS			
	Emotional Support	0.06	0.15
		(0.29)	(0.14)
	Classroom Organization	0.08	0.07
		(0.23)	(0.18)
	Instructional Support	−0.44	0.06
		(0.48)	(0.11)
EF – Hearts and Flowers			
	Emotional Support	−0.00	−0.04
		(0.07)	(0.03)
	Classroom Organization	−0.05	−0.01
		(0.05)	(0.04)
	Instructional Support	−0.18	−0.02
		(0.12)	(0.02)
Math – Applied Problems			
	Emotional Support	−0.02	0.14
		(0.96)	(0.47)
	Classroom Organization	−0.28	0.47
		(0.72)	(0.56)
	Instructional Support	1.41	0.14
		(1.61)	(0.37)
Math – REMA			
	Emotional Support	2.45	−0.54
		(1.69)	(0.92)
	Classroom Organization	1.89	−0.44
		(1.28)	(1.11)
	Instructional Support	3.10	−0.61
		(3.14)	(0.75)

Note: All models control for children's age and baseline score on the requisite outcome, a set of binary indicators for child gender, free or reduced lunch status, home language, class size, and whether the teacher has a master's degree, and a categorical variable indicating the children's race/ethnicity. Robust standard errors (in parentheses) were used to adjust for clustering at the teacher level. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

complete case analysis. For example, we found a small, statistically significant negative linear association between Emotional Support and gains in Hearts and Flowers (see Appendix, Table A5), with a standardized association of −0.14 ($p = 0.02$) in the complete case analysis versus −0.08 ($p = 0.14$) in our multiple imputation findings. Given the number of models we fit for our complete case robustness check and the general agreement of findings across missing data approaches in terms of magnitude and statistical significance, we view these exceptions as likely spurious.

4. Discussion

The purpose of this paper was to replicate and extend a prior analysis examining associations between process quality, as measured by the CLASS, and gains in children's vocabulary, math, and executive function skills in the prekindergarten year in BPS (Weiland et al., 2013). Consistent with the prior study, we found that CLASS scores did not predict gains in children's language across linear and non-linear modeling approaches. While the original study found positive associations between all three CLASS dimensions and children's gains in inhibitory control in higher-quality classrooms, we found null associations between the CLASS domains and our executive function measures across our modeling approaches. In our two-part extension analysis examining math and moderation by child baseline skills, we found no statistically significant associations between any of the CLASS domains and children's gains in math skills (as measured by the REMA) in our quadratic and spline models. We also found no evidence that children's baseline skills moderated associations between any of the CLASS domains and gains in children's skills across the year.

Our null vocabulary findings largely replicate the prior Boston study (Weiland et al., 2013) and the broader preschool literature (Burchinal et al., 2010; Hatfield et al., 2016; Perlman et al., 2016). It may be that process quality as measured by the CLASS does not capture teacher behaviors that are more predictive of preschool children's gains, such as teacher's responsiveness during conversations (i.e., using facilitating peer-to-peer communication, using slow pace to allow children to participate) or the teacher's language complexity (i.e. the length of their sentences) (Justice, Jiang, & Strasser, 2018).

We did not replicate executive function findings from the prior Boston study (Weiland et al., 2013) and from some of the broader literature (Burchinal et al., 2010; Hamre et al., 2014; Mashburn et al., 2008; Perlman et al., 2016), perhaps due to the use of different measures. We used the incongruent trial of Hearts and Flowers to measure inhibitory control, whereas the prior Boston study used the Pencil Tap. The correlation between these two measures when tested in preschool age children is relatively small ($r = 0.28$, $p < 0.001$; Daneri, Sulik, Raver, & Morris, 2018) and there is a paucity of empirical evidence on the extent to which the measures

Table 5
Moderation by child baseline skills (i.e., lowest quartile at baseline).

	Emotional Support	Classroom Organization	Instructional Support
Vocabulary – PPVT			
First Quartile	34.90	20.91	18.64
	(22.25)	(21.61)	(11.64)
Quality	2.77	0.93	−1.37
	(2.27)	(2.46)	(2.08)
Quality × First Quartile	−5.44	−2.94	−4.19
	(3.93)	(3.90)	(3.18)
Inhibitory Control – FDS			
First Quartile	−0.26	−0.14	−0.17
	(1.24)	(1.20)	(0.82)
Quality	0.08	0.03	−0.03
	(0.10)	(0.11)	(0.10)
Quality × First Quartile	0.06	0.04	0.07
	(0.23)	(0.22)	(0.25)
EF – Hearts and Flowers			
First Quartile	−0.06	0.06	−0.10
	(0.47)	(0.53)	(0.27)
Quality	−0.03	−0.03	−0.04
	(0.02)	(0.02)	(0.02)
Quality × First Quartile	−0.00	−0.00	0.01
	(0.09)	(0.10)	(0.08)
Math – Applied Problems			
First Quartile	2.95	1.39	1.12
	(4.89)	(4.51)	(2.50)
Quality	0.13	0.13	0.24
	(0.35)	(0.35)	(0.35)
Quality × First Quartile	−0.41	−0.13	−0.14
	(0.88)	(0.83)	(0.73)
Math – REMA			
First Quartile	4.93	0.22	−1.74
	(7.82)	(7.28)	(4.14)
Quality	0.27	0.33	−0.56
	(0.76)	(0.76)	(0.73)
Quality × First Quartile	−0.58	0.29	1.09
	(1.43)	(1.36)	(1.24)

Note: All models control for children's age and baseline score on the requisite outcome, a set of indicators for child gender, free or reduced lunch status, home language, class size, and whether the teacher has a master's degree, and a categorical variable indicating the children's race/ethnicity. Robust standard errors (in parentheses) were used to adjust for clustering at the teacher level. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

and their scoring rules are comparable. In other work, the Pencil Tap and Hearts and Flowers task have been considered equivalent direct measures of executive function skills (McCoy, 2019). As we noted earlier, the main difference between the two tasks is that the Pencil Tap requires children to inhibit a gross motor behavior in response to an auditive stimulus, whereas the Hearts and Flowers incongruent trial requires children to inhibit the tactile selection of visual stimuli in a digital screen. However, the equivalence of both the cognitive demand and the measurement properties of these tasks have not been demonstrated in the literature. Accordingly, we cannot disentangle whether our findings are due to a lack of sensitivity of the CLASS in measuring classroom factors that promote executive function gains in preschool from differences in EF tasks. More work on measurement of EF tasks is needed to guide applied researchers.

In terms of our math extension work, we found that relations between the CLASS and gains in math were not sensitive to methodological decisions, namely the use of multiple imputation in our regression models, nor to whether the outcome was measured with a more widely used, less sensitive, and more content restrictive measure (Applied Problems) versus a new alternative (REMA). Additional research that uses the REMA – particularly in preschool contexts that use a different math curriculum than Boston – would help illuminate whether our findings extend to other contexts and/or generalize across preschool curriculum approaches.

Regarding the second extension part of our work – moderation by child baseline skills – we had hypothesized that children with lower baseline skills might be more sensitive to classroom process quality. Specifically, we expected initial disadvantages would shrink if children attended high-quality classrooms (versus low-quality classrooms), following the compensatory hypothesis (Sameroff & Chandler, 1975) and empirical findings in the preschool literature to date (Bitler et al., 2014; Bloom & Weiland, 2015; Miller, Farkas, Vandell, & Duncan, 2014). We found no evidence of such moderation for any of our outcomes or CLASS domains, though for statistical power reasons, we were limited to testing for moderation within our linear models. This is an area for further replication and research in other larger samples across varied locations and districts.

Our study has several important limitations. As mentioned, we did not use the Pencil Tap task as the original study did. We used a measure that conceptually taps the same skill – inhibitory control – as the Pencil Tap but we cannot rule out that differences in findings for inhibitory control between the current work and prior Boston work (Weiland et al., 2013) are due to measurement differences. Another limitation is that we did not include literacy or socio-emotional measures and we only capture one element of language development, namely receptive vocabulary. Including a broader range of child skills would have enhanced our study's contribution. Our work is also non-causal. Recent findings in Ecuador with the CLASS have shown larger predictive relations in a study that could identify causality, although in a context with lower structural quality than in Boston (Araujo, Carneiro, Cruz-Aguayo, & Schady, 2016). We found little evidence of sorting into lower and higher quality classrooms by child demographic characteristics or baseline skills (see Appendix, Table A1). Nonetheless, it is possible that our null findings are due to the attenuation of associations by unobservables in our study, such as systematic differences in children's

home environments and immediate communities. In addition, our sample is relatively small and drawn from one district, meaning our power and external validity is more limited than is ideal. Finally, we had a range restriction that did not allow us to test our hypotheses across the expected range of the scales in the CLASS measure. These range restrictions are a near-constant in the field (2011, Burchinal, 2018; Chaudry et al., 2017; Zaslow et al., 2016) but nonetheless, our results could be due to a lack of discrimination, particularly in higher-quality contexts like ours.

Despite these limitations, our findings add to the literature on associations between process quality and preschool children's gains in three ways. First, we build on prior evidence showing that general measures of quality are not strong predictors of language and executive function gains in high-quality contexts. Although we acknowledge that our results are correlational, we minimized differences due to context and curriculum in this replication. Second, our executive function and math results highlight how the outcome measure chosen – even for the same construct – may affect findings in this literature. Third, our null findings for moderation by child baseline skills point to a new direction in this literature that future studies hopefully can replicate to expand our understanding of who benefits from higher classroom quality in preschool.

Our findings also add to the ongoing discussion in the ECE field about the use of general measures of process quality as a component in policy initiatives for supporting children's development (Burchinal, Kainz, & Cai, 2011; Mashburn, 2017; Weiland, 2018; Zaslow, 2011). Such measures do allow for helpful cross-system comparisons and for identifying relative strengths and weaknesses within and across programs (Weiland, 2018). There is also evidence that programs respond to policy-level efforts to increase quality using such measures (Bassok, Dee, & Latham, 2017). But overall, global process quality, as measured by the CLASS, does not consistently predict gains in important domains of early development. It is possible that, in response to increased diversity in preschool settings, programs and teachers are offering a wide range of supports that are directly related with children gains but not captured by general measures of process or structural quality. Since these supports (i.e., offering language resources, use of frequent and systematic testing to inform instruction, fostering or promoting home-literacy practices, among others) rely on the intersection of administrative, pedagogical, and policy decisions, we may need to broaden the scope of our general quality measures to capture these practices.

Accordingly, our findings support calls for a next generation of measurement work in early childhood education (Burchinal, 2018; Weiland, 2018). There are already steps in this direction, including a new language and literacy measure that captures very specific teacher practices in P-3 (Chiang et al., 2017) and new measures that capture children's individual classroom experiences since quality can vary across children in the same classroom (Connor et al., 2009; Sabol, Bohlmann, & Downer, 2018). Similarly, more psychometric development for these measures is needed, in order to demonstrate generalizability properties, measurement invariance, and sensitivity across different levels of general process quality. Given evidence that some preschool curricula are more effective than others (Jenkins et al., 2018; Nguyen, Jenkins, & Auger Whitaker, 2018; Weiland, McCormick, Mattera, Maier, & Morris, 2018), new measures that take curricula into account may also be a fruitful area for new research.

5. Declaration of interest

None.

CRediT authorship contribution statement

Paola Guerrero-Rosada: Conceptualization, Data curation, Validation, Formal analysis, Writing - original draft, Visualization. **Christina Weiland:** Conceptualization, Methodology, Validation, Formal analysis, Resources, Writing - original draft, Supervision, Project administration, Funding acquisition. **Meghan McCormick:** Conceptualization, Validation, Investigation, Resources, Data curation, Writing - review & editing, Supervision, Project administration, Funding acquisition. **JoAnn Hsueh:** Conceptualization, Investigation, Resources, Supervision, Funding acquisition, Project administration. **Jason Sachs:** Conceptualization, Investigation, Supervision, Funding acquisition, Project administration. **Catherine Snow:** Conceptualization, Investigation, Resources, Supervision, Funding acquisition, Project administration. **Michelle Maier:** Conceptualization, Investigation, Resources, Data curation, Supervision, Funding acquisition.

Appendix A. Robustness Checks

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.ecresq.2020.07.009.

References

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., & Schady, N. (2016). Teacher quality and learning outcomes in kindergarten. *The Quarterly Journal of Economics*, *131*(3), 1415–1453. http://dx.doi.org/10.1093/qje/qjw016

Barrueco, S., López, M., Ong, C., & Lozano, P. (2012). *Assessing Spanish–English bilingual preschoolers: A guide to best approaches and measures*. Paul H Brookes Publishing.

Bassok, D., Dee, T., & Latham, S. (2017). *The effects of accountability incentives in early childhood education (No. w23859)*. http://dx.doi.org/10.3386/w23859

Bitler, M., Hoynes, H., & Domina, T. (2014). *Experimental evidence on distributional effects of Head Start (No. w20434)*. http://dx.doi.org/10.3386/w20434

Bloom, H. S., & Weiland, C. (2015). Quantifying variation in head start effects on young children's cognitive and socio-emotional skills using data from the National Head Start Impact Study. *SSRN Electronic Journal*, http://dx.doi.org/10.2139/ssrn.2594430

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

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

Burchinal, M., Kainz, K., & Cai, Y. (2011). How well do our measures of quality predict child outcomes? In *Quality Measurement in Early Childhood Settings*. pp. 11–31. Baltimore, MD, US: Paul H. Brookes Pub. Co.

Burchinal, M., Vandergrift, N., Pianta, R., & Mashburn, A. (2010). Threshold analysis of association between child care quality and child outcomes for low-income children in pre-kindergarten programs. *Early Childhood Research Quarterly*, *25*(2), 166–176. http://dx.doi.org/10.1016/j.ecresq.2009.10.004

Burchinal, M., Vernon-Feagans, L., Vitiello, V., & Greenberg, M. (2014). Thresholds in the association between child care quality and child outcomes in rural preschool children. *Early Childhood Research Quarterly*, *29*(1), 41–51. http://dx.doi.org/10.1016/j.ecresq.2013.09.004

Camerota, M., Willoughby, M. T., & Blair, C. B. (2019). Speed and accuracy on the hearts and flowers task interact to predict child outcomes. *Psychological Assessment*, *31*(8), 995–1005. http://dx.doi.org/10.1037/pas0000725

Chaudry, A., Morrissey, T., Weiland, C., & Yoshikawa, H. (2017). *Cradle to kindergarten: A new plan to combat inequality*. New York: Russell Sage Foundation.

Chiang, H., Walsh, E., Shanahan, T., Gentile, C., Maccarone, A., Waits, T., . . . & Rikoon, S. (2017). *An exploration of instructional practices that foster language development and comprehension: Evidence from prekindergarten through grade 3 in Title I schools (NCEE 2017-4024)*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Clements, D. H., & Sarama, J. (2008). Experimental evaluation of the effects of a research-based preschool mathematics curriculum. *American Educational Research Journal*, *45*(2), 443–494. http://dx.doi.org/10.3102/0002831207312908

Clements, D. H., & Sarama, J. (2011). Early childhood mathematics intervention. *Science*, *333*(6045), 968–970. http://dx.doi.org/10.1126/science.1204537

Clements, D. H., Sarama, J. H., & Liu, X. H. (2008). Development of a measure of early mathematics achievement using the Rasch model: The research-based

early maths assessment. *Educational Psychology*, *28*(4), 457–482. http://dx.doi.org/10.1080/01443410701777272
Connor, C. M., Morrison, F. J., Fishman, B. J., Ponitz, C. C., Glasney, S., Underwood, P. S., . . . & Schatschneider, C. (2009). The ISI classroom observation system: Examining the literacy instruction provided to individual students. *Educational Researcher*, *38*(2), 85–99. http://dx.doi.org/10.3102/0013189X09332373
Curby, T. W., Johnson, P., Mashburn, A. J., & Carlis, L. (2016). Live versus video observations: Comparing the reliability and validity of two methods of assessing classroom quality. *Journal of Psychoeducational Assessment*, *34*(8), 765–781. http://dx.doi.org/10.1177/0734282915627115
Daneri, M. P., Sulik, M. J., Raver, C. C., & Morris, P. A. (2018). Observers' reports of self-regulation: Measurement invariance across sex, low-income status, and race/ethnicity. *Journal of Applied Developmental Psychology*, *55*, 14–23. http://dx.doi.org/10.1016/j.appdev.2017.02.001
Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia*, *44*(11), 2037–2078.
Diamond, A., & Taylor, C. (1996). Development of an aspect of executive control: Development of the abilities to remember what I said and to "Do as I say, not as I do". *Developmental Psychobiology*, *29*(4), 315–334.
Diamond, A., Barnett, W. S., Thomas, J., & Munro, S. (2007). Preschool program improves cognitive control. *Science*, *318*(5855), 1387–1388. http://dx.doi.org/10.1126/science.1151148
Duncan, S. E., & De Avila, E. A. (1998). *Pre-LAS 2000*. Monterey, CA: CTB/McGraw-Hill.
Duncan, G. J., Engel, M., Claessens, A., & Dowsett, C. J. (2014). Replication and robustness in developmental research. *Developmental Psychology*, *50*(11), 2417–2425. http://dx.doi.org/10.1037/a0037996
Dunn, L. M., & Dunn, D. M. (2007). *PPVT-4: Peabody picture vocabulary test. Pearson assessments*.
Early, D. M., Maxwell, K. L., Ponder, B. D., & Pan, Y. (2017). Improving teacher-child interactions: A randomized controlled trial of making the most of classroom interactions and my teaching partner professional development models. *Early Childhood Research Quarterly*, *38*, 57–70. http://dx.doi.org/10.1016/j.ecresq.2016.08.005
Friedman-Krauss, A., Barnett, W. S., Weisenfeld, G., Kasmin, R., DiCrecchio, N., & Horowitz, M. (2018). *The state of preschool 2017 (state preschool yearbook)* Retrieved from. Rutgers, New York: The National Institute for Early Education Research. http://nieer.org/state-preschool-yearbooks/yearbook2017
Giofrè, D., Stoppa, E., Ferioli, P., Pezzuti, L., & Cornoldi, C. (2016). Forward and backward digit span difficulties in children with specific learning disorder. *Journal of Clinical and Experimental Neuropsychology*, *38*(4), 478–486. http://dx.doi.org/10.1080/13803395.2015.1125454
Gormley, W. T., Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal Pre-K on cognitive development. *Developmental Psychology*, *41*(6), 872–884. http://dx.doi.org/10.1037/0012-1649.41.6.872
Graham, J. W. (2012). *Missing data: Analysis and design*. http://dx.doi.org/10.1007/978-1-4614-4018-5
Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*(1), 549–576. http://dx.doi.org/10.1146/annurev.psych.58.110405.085530
Hamre, B., Hatfield, B., Pianta, R., & Jamil, F. (2014). Evidence for general and domain-specific elements of teacher–child interactions: Associations with preschool children's development. *Child Development*, *85*(3), 1257–1274. http://dx.doi.org/10.1111/cdev.12184
Hamre, B., Pianta, R., Burchinal, M., Field, S., LoCasale-Crouch, J., Downer, J. T., . . . & Scott-Little, C. (2012). A course on effective teacher–child interactions: Effects on teacher beliefs, knowledge, and observed practice. *American Educational Research Journal*, *49*(1), 88–123. http://dx.doi.org/10.3102/0002831211434596
Hatfield, B., Burchinal, M., Pianta, R., & Sideris, J. (2016). Thresholds in the association between quality of teacher–child interactions and preschool children's school readiness skills. *Early Childhood Research Quarterly*, *36*, 561–571. http://dx.doi.org/10.1016/j.ecresq.2015.09.005
Hilbert, S., Nakagawa, T. T., Bindl, M., & Bühner, M. (2014). The spatial Stroop effect: A comparison of color-word and position-word interference. *Psychonomic Bulletin & Review*, *21*(6), 1509–1515. http://dx.doi.org/10.3758/s13423-014-0631-4
Hillemeier, M. M., Morgan, P. L., Farkas, G., & Maczuga, S. A. (2013). Quality disparities in child care for at-risk children: comparing head start and non-head start settings. *Maternal and Child Health Journal*, *17*(1), 180–188. http://dx.doi.org/10.1007/s10995-012-0961-7
Hsueh, J. (2016). *Boston P-3: Identifying malleable factors for promoting student success*. Institute of Education Sciences - Funded Research Grants and Contracts. https://ies.ed.gov/funding/grantsearch/details.asp?ID=1770
Institute of Education Sciences (IES), & U.S. Department of Education and National Science Foundation (NSF). (2018). *Companion Guidelines on Replication & Reproducibility in Education Research* Retrieved from IES website:. https://ies.ed.gov/pdf/CompanionGuidelinesReplicationReproducibility.pdf
Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, Article e124 http://dx.doi.org/10.1371/journal.pmed.0020124
Jenkins, J. M., Duncan, G. J., Auger, A., Bitler, M., Domina, T., & Burchinal, M. (2018). Boosting school readiness: Should preschool teachers target skills or the whole child? *Economics of Education Review*, *65*, 107–125. http://dx.doi.org/10.1016/j.econedurev.2018.05.001
Justice, L. M., Jiang, H., & Strasser, K. (2018). Linguistic environment of preschool classrooms: What dimensions support children's language growth? *Early Childhood Research Quarterly*, *42*, 79–92. http://dx.doi.org/10.1016/j.ecresq.2017.09.003
Keys, T. D., Farkas, G., Burchinal, M. R., Duncan, G. J., Vandell, D. L., Li, W., . . . & Howes, C. (2013). Preschool center quality and school readiness: Quality effects and variation by demographic and child characteristics. *Child Development*, *84*(4), 1171–1190.
Leyva, D., Weiland, C., Barata, M., Yoshikawa, H., Snow, C., Treviño, E., . . . & Rolla, A. (2015). Teacher–child interactions in Chile and their associations with prekindergarten outcomes. *Child Development*, *86*(3), 781–799. http://dx.doi.org/10.1111/cdev.12342
Lima, A., & Melnik, M. (2014). *Boston: Measuring diversity in a changing city* Retrieved from City of Boston website:. http://www.bostonplans.org/getattachment/32e9b68a-ce1b-41c7-808c-0395cb4f4d19
Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, *43*(6), 304–316. http://dx.doi.org/10.3102/0013189X14545513
Marsh, L. C., & Cormier, D. R. (2002). *Spline regression models*. London, UK: Sage Publications.
Mashburn, A. J. (2017). Evaluating the validity of classroom observations in the head start designation renewal system. *Educational Psychologist*, *52*(1), 38–49. http://dx.doi.org/10.1080/00461520.2016.1207539
Mashburn, A. J., Downer, J. T., Hamre, B. K., Justice, L. M., & Pianta, R. C. (2010). Consultation for teachers and children's language and literacy development during pre-kindergarten. *Applied Developmental Science*, *14*(4), 179–196. http://dx.doi.org/10.1080/10888691.2010.516187
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*(3), 732–749. http://dx.doi.org/10.1111/j.1467-8624.2008.01154.x
McCoy, D. C. (2019). Measuring young children's executive function and self-regulation in classrooms and other real-world settings. *Clinical Child and Family Psychology Review*, *22*(1), 63–74. http://dx.doi.org/10.1007/s10567-019-00285-1
Miller, E. B., Farkas, G., & Duncan, G. J. (2016). Does head start differentially benefit children with risks targeted by the program's service model? *Early Childhood Research Quarterly*, *34*, 1–12. http://dx.doi.org/10.1016/j.ecresq.2015.08.001
Miller, E. B., Farkas, G., Vandell, D. L., & Duncan, G. J. (2014). Do the effects of head start vary by parental preacademic stimulation? *Child Development*, *85*(4), 1385–1400. http://dx.doi.org/10.1111/cdev.12233
Muller, U., Kerns, K. A., & Konkin, K. (2012). Test–retest reliability and practice effects of executive function tasks in preschool children. *The Clinical Neuropsychologist*, *26*(2), 271–287. http://dx.doi.org/10.1080/13854046.2011.645558
Nguyen, T., Jenkins, J. M., & Auger Whitaker, A. (2018). Are content-specific curricula differentially effective in head start or state prekindergarten classrooms? *AERA Open*, *4*(2), Article 233285841878428 http://dx.doi.org/10.1177/2332858418784283
Nguyen, T., Watts, T. W., Duncan, G. J., Clements, D. H., Sarama, J. S., Wolfe, C., . . . & Spitler, M. E. (2016). Which preschool mathematics competencies are most predictive of fifth grade achievement? *Early Childhood Research Quarterly*, *36*, 550–560. http://dx.doi.org/10.1016/j.ecresq.2016.02.003
Pakarinen, E., Lerkkanen, M., Poikkeus, A., Siekkinen, M., & Nurmi, J. (2011). Kindergarten teachers adjust their teaching practices in accordance with children's academic pre-skills. *Educational Psychology*, *31*(1), 37–53. http://dx.doi.org/10.1080/01443410.2010.517906
Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*(6), 531–536. http://dx.doi.org/10.1177/1745691612463401
Perlman, M., Falenchuk, O., Fletcher, B., McMullen, E., Beyene, J., & Shah, P. S. (2016). A systematic review and meta-analysis of a measure of staff/child interaction quality (the classroom assessment scoring system) in early childhood education and care settings and child outcomes. *PLoS One*, *11*(12), Article e0167660 http://dx.doi.org/10.1371/journal.pone.0167660
Phillips, D. A., Lipsey, M., Dodge, K., Haskins, R., Bassok, D., Burchinal, M., . . . & Weiland, C. (2017). *Puzzling it out: The current state of scientific knowledge on pre-kindergarten effects* Retrieved from. https://www.brookings.edu/wp-content/uploads/2017/04/duke_prekstudy_final_4-4-17_hires.pdf
Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, *38*(2), 109–119. http://dx.doi.org/10.3102/0013189X09332374
Pianta, R., Hamre, B., Downer, J., Burchinal, M., Williford, A., LoCasale-Crouch, J., . . . & Scott- Little, C. (2017). Early childhood professional development: Coaching and coursework effects on indicators of children's school readiness. *Early Education and Development*, *28*(8), 956–975. http://dx.doi.org/10.1080/10409289.2017.1319783
Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system™: Manual K-3*. Baltimore, MD, US: Paul H Brookes Publishing.
Puma, M., Bell, S., Cook, R., Heid, C., & U.S Department of Health and Human Services. (2010). *Head start impact study final report*. Office of Planning, Research and Evaluation, Administration for Children and Families, US Department of Health and Human Services.

Raver, C. C., Jones, S. M., Li-Grining, C., Zhai, F., Bub, K., & Pressler, E. (2011). CSRP's impact on low-income preschoolers' preacademic skills: Self-regulation as a mediating mechanism: CSRP's impact on low-income preschoolers' preacademic skills. *Child Development*, *82*(1), 362–378. http://dx.doi.org/10.1111/j.1467-8624.2010.01561.x

Sabol, T. J., Bohlmann, N. L., & Downer, J. T. (2018). Low-income ethnically diverse children's engagement as a predictor of school readiness above preschool classroom quality. *Child Development*, *89*(2), 556–576. http://dx.doi.org/10.1111/cdev.12832

Sameroff, A. J., & Chandler, M. J. (1975). Reproductive risk and the continuum of caretaker casualty. In F. D. Horowitz (Ed.), *Review of child development research* (Vol. 4). Chicago: University of Chicago Press.

Sarama, J., & Clements, D. H. (2004). Building blocks for early childhood mathematics. *Early Childhood Research Quarterly*, *19*(1), 181–189. http://dx.doi.org/10.1016/j.ecresq.2004.01.014

Schauer, J. M. (2018). *Statistical methods for assessing replication: A meta-analytic framework* Retrieved from. ProQuest Dissertations & Theses Global., Article 2164811196 https://proxy.lib.umich.edu/login?url=https://search-proquest-com.proxy.lib.umich.edu/docview/2164811196?accountid=14667

Schickedanz, J., & Dickinson, D. (2004). *Opening the world of learning: A comprehensive early literacy program*. Pearson Early Learning.

U.S. Department of Health and Human Services. (2018). *Use of Classroom Assessment Scoring System (CLASS®) in Head Start* Retrieved February 3, 2019, from. https://eclkc.ohs.acf.hhs.gov/designation-renewal-system/article/use-classroom-assessment-scoring-system-classr-head-start

van Ginkel, J. R., Linting, M., Rippe, R. C. A., & van der Voort, A. (2020). Rebutting existing misconceptions about multiple imputation as a method for handling missing data. *Journal of Personality Assessment*, *102*(3), 297–308. http://dx.doi.org/10.1080/00223891.2018.1530680

von Hippel, P. T. (2009). How to impute interactions, squares, and other transformed variables. *Sociological Methodology*, *39*(1), 265–291. http://dx.doi.org/10.1111/j.1467-9531.2009.01215.x

Wechsler, D. (1974). . *Manual for the Wechsler intelligence scale for children, revised* (Vols. 1–vii) New York, NY: Psychological Corp.

Weiland, C. (2018). Commentary: Pivoting to the "how": Moving preschool policy, practice, and research forward. *Early Childhood Research Quarterly*, *45*, 188–192. http://dx.doi.org/10.1016/j.ecresq.2018.02.017

Weiland, C., McCormick, M., Mattera, S., Maier, M., & Morris, P. (2018). Preschool curricula and professional development features for getting to high-quality implementation at scale: A comparative review across five trials. *AERA Open*, *4*(1), Article 233285841875773 http://dx.doi.org/10.1177/2332858418757735

Weiland, C., Ulvestad, K., Sachs, J., & Yoshikawa, H. (2013). Associations between classroom quality and children's vocabulary and executive function skills in an urban public prekindergarten program. *Early Childhood Research Quarterly*, *28*(2), 199–209. http://dx.doi.org/10.1016/j.ecresq.2012.12.002

Weiland, C., Wolfe, C. B., Hurwitz, M. D., Clements, D. H., Sarama, J. H., & Yoshikawa, H. (2012). Early mathematics assessment: Validation of the short form of a prekindergarten and kindergarten mathematics measure. *Educational Psychology*, *32*(3), 311–333. http://dx.doi.org/10.1080/01443410.2011.654190

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817–838. http://dx.doi.org/10.2307/1912934

Whitehurst, G., & Klein, E. (2015). *Do we already have universal preschool?* Washington, DC: Economic Studies at Brookings.

Wong, V. C., Cook, T. D., Barnett, W. S., & Jung, K. (2008). An effectiveness-based evaluation of five state pre-kindergarten programs. *Journal of Policy Analysis and Management*, *27*(1), 122–154. http://dx.doi.org/10.1002/pam.20310

Woodcock, R. W., Mather, N., McGrew, K. S., & Wendling, B. J. (2001). *Woodcock–Johnson III tests of cognitive abilities*. Itasca, IL: Riverside Publishing Company.

Woodcock, R. W., Munoz-Sandoval, A. F., Ruef, M. L., & Alvarado, C. G. (2005). *Bateria III Woodcock–Munoz: Pruebas de habilidades cognitivas*. Riverside Publishing Company.

Wright, A., & Diamond, A. (2014). An effect of inhibitory load in children while keeping working memory load constant. *Frontiers in Psychology*, *5* http://dx.doi.org/10.3389/fpsyg.2014.00213

Yoshikawa, H., Weiland, C., Brooks-Gunn, J., Burchinal, M., Espinosa, L. M., Gormley, W. T., . . . & Zaslow, M. (2013). *Investing in our future: The evidence base on preschool education* Retrieved from:. https://www.srcd.org/policy-media/policy-updates/meetings-briefings/investing-our-future-evidence-base-preschool

Zaslow, M. (Ed.). (2011). *Quality measurement in early childhood settings*. Baltimore, Md: Paul H. Brookes Pub. Co.

Zaslow, M., Anderson, R., Redd, Z., Wessel, J., Daneri, P., Green, K., . . . & Martinez-Beck, I. (2016). Quality thresholds, features, and dosage in early care and education. *Monographs of the Society for Research in Child Development*, *81*(2), 7–26. http://dx.doi.org/10.1111/mono.12236

Zaslow, M., Anderson, R., Redd, Z., Wessel, J., Tarullo, L., & Burchinal, M. (2011). *Quality dosage, thresholds, and features in early childhood settings: A review of the literature. (No. 06671.310)*. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.: OPRE.