Early Childhood Research Quarterly 28 (2013) 199–209

Contents lists available at SciVerse ScienceDirect

Early Childhood Research Quarterly

Associations between classroom quality and children's vocabulary and executive function skills in an urban public prekindergarten program☆

Christina Weiland[a,*,1], Kchersti Ulvestad[a], Jason Sachs[b], Hirokazu Yoshikawa[a,1]

[a] Harvard Graduate School of Education, United States
[b] Boston Public Schools, United States

ARTICLE INFO

Article history:
Received 7 November 2011
Received in revised form
29 November 2012
Accepted 5 December 2012

Keywords:
Classroom quality
Receptive vocabulary
Executive function

ABSTRACT

Despite evidence that high-quality preschool programs have substantial, long-lasting impacts on young children's developmental outcomes, associations between preschool quality measures and children's cognitive outcomes within preschool programs are generally small or null. Using data from a large urban prekindergarten program, we examined associations between children's receptive vocabulary and executive function skills and several indicators of classroom quality. Ours is the first such study within a program that has been shown to have small-to-large causal impacts on children's language, literacy, mathematics, executive function, and emotional development outcomes. Consistent with prior literature, we found small or null associations between quality predictors and children's outcomes and we found that some of these relationships were curvilinear. Findings are discussed in light of several hypotheses in the literature regarding the general pattern of small or null associations, including the psychometrics of commonly used quality measures and possible range restriction of quality indicators.

Early care settings are theorized to be an important proximal context of children's development (Bronfrenbrenner & Morris, 1998). High-quality settings are replete with the kind of learning opportunities that theory suggests promote positive development and prepare children for school, including exposure to new vocabulary and early mathematics concepts, positive peer interactions, and other rich learning opportunities (Dickinson & Smith, 2001; Neuman & Carta, 2011; NICHD & Duncan, 2003). In the empirical literature, there is considerable support for the developmental importance of high-quality early care settings. Several studies have found that high-quality preschool experiences can have long-lasting positive effects on children's life outcomes (Campbell & Ramey, 1994; Schweinhart et al., 2005). More recent causal evaluations have found small-to-moderate impacts of prekindergarten programs on children's short-term developmental outcomes (Gormley, Gayer, Phillips, & Dawson, 2005; Hustedt, Barnett, Jung, & Thomas, 2007; Hustedt, Barnett, Jung, & Goetze, 2009; Weiland & Yoshikawa, in press; Wong, Cook, Barnett, & Jung, 2008).

However, within programs, there is a puzzle in the empirical literature. That is, the relationships between various quality indicators, including teacher qualifications and observational quality measures, and gains in children's outcomes are generally weak (Zaslow et al., 2010; Zaslow, Martinez-Beck, Tout, & Halle, 2011). Scholars, particularly Burchinal, Kainz, and Cai (2011), have articulated several hypotheses as to why. First, it could be that despite theory and causal evaluations of preschool programs such as Abecedarian, Perry, and state prekindergarten programs that suggest otherwise, early care setting quality does not have much effect on child outcomes. Some scholars such as Blau (1999) have argued that this is the case. Second, current measures of quality – including teacher characteristics, program standards, and observational measures – may not be adequately measuring the construct of classroom quality. For example, recent work on the ECERS-R, a standard observational quality measure, has found that it has several psychometric problems, including disordered items, that may affect the results of analyses of this measure (Cassidy, Hestenes, Hegde, Hestenes, & Mims, 2005; Gordon, Fujimoto, Kaestner, Korenman, & Abner, 2010). Third, measures of specific aspects of classroom quality, such as the quality of mathematics, literacy, and language instructions, may be needed instead of more commonly used general structural, emotional, and instructional quality measures. Fourth, the majority of studies rate programs as mediocre or low on instructional quality, the dimension of quality

☆ Special thanks to participating families, teachers, principals, BPS NAEYC coordinator Karen Silver, early childhood coaches, the Wellesley Centers for Women, and district staff.

* Corresponding author at: Harvard Graduate School of Education, 14 Appian Way, Room 705, Cambridge, MA 02139, United States. Tel.: +1 617 276 4773; fax: +1 617 496 5191.
E-mail address: Christina_weiland@mail.harvard.edu (C. Weiland).

[1] Christina Weiland and Hirokazu Yoshikawa's work on this study was funded by the Institute for Education Sciences.

0885-2006/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ecresq.2012.12.002

that may matter most in supporting growth in children's academic outcomes. Range restriction, thus, may also be a contributing reason that detected relationships are generally weak. Finally, the findings in this literature come from non-experimental studies and as such are subject to the usual host of concerns regarding selection and omitted variables bias, although recent work with a quasi-experimental design found similarly weak relationships between observational quality measures and children's mathematics, language, and literacy outcomes (Auger, Farkas, Duncan, Burchinal, & Vandell, 2012).

Whatever the reason(s), the pattern of weak relationships between child outcomes and the current indicators of quality is generally consistent across different types of early care settings and across different types of model specifications, including linear, non-linear, and threshold/spline regression models and across multiple regression, change score, and residualized change approaches (Burchinal, Vandergrift, Pianta, & Mashburn, 2010; Burchinal, Kainz, et al., 2011; NICHD & Duncan, 2003; Zaslow et al., 2010, 2011). In the current study, we contribute to this literature in four main ways. First, using linear and quadratic model specifications, we explored whether classroom quality rating scales from a large publicly funded prekindergarten program in the Boston Public Schools (BPS) predicted gains in children's receptive vocabulary outcomes. Importantly, the BPS program has higher-than-average instructional quality ratings compared to other studies in this literature, meaning we were able to address questions regarding whether previously detected weak relationships were due to range restrictions in instructional quality measures. Second, we examined the relationship between quality indicators and executive function, a developmentally important domain of school readiness (Blair & Razza, 2007) that has not been included in prior studies of the relationship between classroom quality and children's outcomes (Zaslow et al., 2010). Third, ours is the first study in this literature to examine the quality and child outcomes link within a proven prekindergarten program – a program in which the majority of teachers had masters degrees and implemented consistent, proven mathematics, language and literacy curricula. A recent regression discontinuity design evaluation found that the BPS program has moderate-to-large impacts on children's language, literacy and mathematics outcomes and small impacts on children's executive function and emotional development outcomes (Weiland & Yoshikawa, in press). As such, the present study represents an intersection of the literature on the causal effects of preschool on children's outcomes and the non-experimental literature on the associations between program quality and children's outcomes. It also provides additional contextual information on the BPS prekindergarten program, as the Weiland and Yoshikawa (in press) study did not include data from the observational classroom quality measures used in the present study.

1. Why classroom quality may impact young children's language and executive function

From a theoretical perspective, higher early childhood classroom quality – as characterized by rich learning opportunities, positive peer interactions, positive student–teacher relationships, adequate staff–child ratios, adequate learning materials, and safe, clean facilities – is hypothesized to promote positive child development, including improved child language and executive function skills. In terms of mechanisms, high-quality classrooms build student language by providing ample oral language opportunities and exposing children to new vocabulary (Dickinson & Smith, 2001; Neuman & Carta, 2011; NICHD & Duncan, 2003). Some empirical literature finds support for the theorized links between children's language development and classroom quality. For example, several studies have found that children in higher-quality programs outscore their peers in lower-quality programs on different measures of expressive and receptive language, although associations are in the small range (Burchinal et al., 2010; Howes et al., 2008; Mashburn et al., 2008). Within those same studies, however, associations are not consistent across different measures of quality; that is, some quality scales predict gains in child language outcomes and some do not. For example, Mashburn et al. (2008) found that a structural quality measure and an emotional climate measure did not predict children's receptive vocabulary scores, although an instructional quality measure did. Interestingly, but not surprisingly, associations tend to be larger between children's cognitive outcomes and instructional/academic quality measures than between children's cognitive outcomes and emotional climate measures (Burchinal et al., 2008, 2010; Howes et al., 2008; Mashburn et al., 2008).

In terms of executive function (EF) – a set of cognitive processes integral to the emerging self-regulation of behavior and the development of social and cognitive competence in young children (Blair, 2002) – some scholars have proposed that higher-quality classrooms may build children's EF by supporting child activity choice and reflection (Bodrova & Leong, 2006). Others have suggested that better classroom management may promote children's executive function skills, as children in better-managed classrooms may internalize modeled regulatory behaviors and thus show improved self-behavior management (Rimm-Kaufman, Curby, Grimm, Nathanson, & Brock, 2009). The quality of the language and literacy environment may matter as well for EF; a higher-quality language and literacy environment could improve children's receptive vocabulary, which could build children's executive function skills by enhancing children's outer and then inner speech. Improved self-talk in turn may improve executive functions, as children become better able to plan and monitor their behavior (Fuhs & Day, 2011; Zakin, 2007).

The relationship between classroom quality and executive function generally has been underexplored in the empirical preschool literature (Zaslow et al., 2010). This is an important gap in the literature, as executive function is developing rapidly during the preschool years (Welsh, Nix, Blair, Bierman, & Nelson, 2010; Zelazo & Müller, 2002) and it appears to underlie later school success (Blair & Razza, 2007; Howse, Lange, Farran, & Boyles, 2003; McClelland, Morrison, & Holmes, 2000; McClelland, Acock, & Morrison, 2006; McClelland et al., 2007; Welsh et al., 2010). Only one previous study to our knowledge has examined links between classroom quality and executive function (Rimm-Kaufman et al., 2009). That study found that higher classroom organization was positively associated with children's cognitive self-control but that the higher instructional support was negatively associated with children's cognitive self-control, controlling for children's fall self-regulation skills. Higher emotional support was not associated with children's cognitive self-control. Another study found that an executive function intervention increased both classroom quality and children's executive function skills (Barnett et al., 2008) but that study did not examine the relationship between classroom quality and children's executive function.

1.1. Modeling the relationship between child outcomes and classroom quality

The relationship between child outcomes and classroom quality, as measured by continuous observational quality measures, has most commonly been modeled as linear (Burchinal et al., 2008; Mashburn et al., 2008; Zaslow et al., 2010). As questions have been posed about the relatively small associations between classroom quality and child outcomes within programs, researchers increasingly have begun to explore whether these relationships may be

non-linear. Theoretically, classroom quality may not make a difference in children's outcomes until a certain level of quality is reached (as would be evidenced by a steeper, positive slope in the higher quality range compared to the slope in the lower quality range). Conversely, there may be diminishing returns to children's outcomes at higher levels of quality (as would be evidenced by a flatter positive or perhaps even a negative slope in the higher quality range compared to the slope in the lower quality range). Burchinal, Kainz, et al. (2011) tested for quadratic relationships between child language, literacy, mathematics and socio-emotional outcomes and several commonly used quality measures. They found statistically significant quadratic relationships for 8 out of 28 examined outcome/quality combinations, with quality generally more strongly related to outcomes when quality was in the higher range. Other previous studies that have examined nonlinear relationships between observational quality measures and child outcomes have concluded that these relationships are linear (NICHD ECCRN, 2006; NICHD & Duncan, 2003; Peisner-Feinberg et al., 2001).

More recently, researchers have used spline regression models to examine whether relationships between observational quality ratings measures and child outcomes are stronger in certain quality ranges (Zaslow et al., 2010). Spline regression models are essentially models in which the fitted line is permitted to change slope but not intercept at a particular knot or threshold (Marsh & Cormier, 2002). As Zaslow et al. (2010) explain, these models extend prior work that examined nonlinear relationships in this literature. By permitting tests of whether relationships are stronger in certain quality ranges, they move beyond reporting whether the functional form of these relationships is linear or nonlinear. Spline regression models also potentially have different policy implications compared to prior nonlinear modeling approaches in this literature. Perhaps most compellingly, they could potentially answer policy-relevant questions about what level of quality is "good enough" (Burchinal et al., 2010). If there are thresholds above which there are limited returns to quality investments, then resources might be best spent raising programs up to that threshold but not beyond. Burchinal et al. (2010) used spline regression models to examine whether relationships between two classroom quality measures and child language, literacy, mathematics and socio-emotional outcomes were stronger in certain quality ranges. They found that the slope of the relationship between classroom quality and child outcomes generally was a stronger predictor in higher versus lower ranges of quality. Nonetheless, with this approach too, associations were generally in the small range, meaning that the weak relationships between child outcomes and program quality do not appear to be a function solely of modeling approach.

1.2. Current study

In the present study, we add to the prekindergarten literature by examining the relationship between indicators of quality and children's outcomes within a public prekindergarten program. Specifically, we explored the following research question: Is higher classroom quality, as measured by several commonly used early childhood classroom quality measures, associated with higher end-of-prekindergarten child receptive vocabulary and executive function skills?

2. Method

2.1. Participants and setting

The study sample included 414 children attending the Boston Public Schools public prekindergarten program in 2009–2010. All programs were located in the public schools, and all teachers were BPS employees with the same compensation and requirements as their K-12 peers. Study children were nested within 83 classrooms in 46 schools, an average of about 5 children per classroom and 1.8 classrooms per school. The prekindergarten program was open to any 4-year-old in the district; there were no income requirements or other criteria for enrollment as there are in many prekindergarten programs (Barnett et al., 2010). Sample children represented a subset of the 1049 children who agreed to participate in an evaluation of the 4-year-old prekindergarten program in fall 2009 (Weiland & Yoshikawa, in press). In spring 2010, 414 children tested in fall 2009 were randomly selected to be tested again, as part of the district's regular bi-annual program monitoring.

Study children were diverse in their background characteristics. Overall, 68% of sample children received free/reduced lunch, 13% were classified as having a special need, and 50% were male. In terms of language, 27% spoke Spanish at home, 52% spoke English at home, and the remaining 21% spoke a language other than Spanish or English at home. They were also racially/ethnically diverse: 43% were Hispanic, 28% were Black, 16% were White, 11% were Asian and 3% were mixed/other. Participating children were approximately representative of the district's prekindergarten population, although there were statistically significant differences between the overall district's 4-year-old population and the sample population on eight out of twenty of the available background characteristics: Asian (11% participating vs. 7% not participating; $p < .05$), English home language (52% participating vs. 61% not participating; $p < .001$), home language other than English or Spanish (21% participating vs. 16% not participating; $p < .01$), free/reduced lunch status (68% participating vs. 50% not participating; $p < .001$), enrollment/withdrawal from more than one school during the year (12% participating vs. 23% not participating; $p < .001$), east attendance zone (45% participating vs. 38% not participating; $p < .01$), west attendance zone (26% participating vs. 32% not participating; $p < .05$), and previous attendance at a private preschool (29% participating vs. 35% not participating; $p < .05$).

All study teachers had at least a bachelor's degree and 83% held a master's degree. In terms of degree concentrations, 36% of teachers held a bachelor's degree in early childhood education and 43% held a master's degree in early childhood education. Study teachers were also relatively experienced: 7% percent had 0–3 years of teaching experience, 11% had 3–5 years of teaching experience, 26% had 5–10 years of teaching experience, and 57% had over 10 years of teaching experience. In terms of class size, 46% of classrooms had 20 students or less. Participating prekindergarten teachers and classrooms were approximately representative of all prekindergarten teachers and classrooms in the district. Out of 23 tested characteristics, there were statistically significant differences between participating and non-participating classrooms on two characteristics: class size less than twenty (46% participating vs. 87% not participating; $p < .001$) and bachelor's degree in early childhood education (36% participating vs. 18% not participating; $p < .05$).

In the study year, a literacy curriculum, *Opening the World of Learning* (OWL; Schickedanz & Dickinson, 2005), and a mathematics curriculum, *Building Blocks* (Clements & Sarama, 2007a), were in place system wide in prekindergarten classrooms and had been in place since the 2007–2008 school year. Both curricula have shown positive results in other contexts (Ashe, Reed, Dickinson, Morse, & Wilson, 2009; Clements & Sarama, 2007b), although results are mixed for the OWL (Dickinson et al., 2011). Weiland and Yoshikawa (in press) found that in the year prior to the present study year, curricula were moderately to highly well implemented across the district. While the OWL directly targets vocabulary, neither the OWL nor Building Blocks specifically targets executive function skills . However, emerging evidence suggests that improving children's language, literacy, and mathematics skills may also improve their

executive function skills (Fuhs & Day, 2011; Weiland & Yoshikawa, in press; Welsh et al., 2010).

2.2. Procedures

Child assessments were conducted in one-on-one pull-out sessions in the child's school with trained child assessors. In the fall session, child assessors had to prove reliability on the battery of tests and show good rapport/child management skills in both simulated and real testing situations. On average, the battery of tests took approximately 45–50 min to administer and children were assessed on ten tests. Assessors were instructed to test children in one session if possible but to divide the session into smaller segments if children showed signs of fatigue. Because of the session length, we randomly varied the order of the tests to limit the possibility of systematically biasing results due to child fatigue.

A different study team assessed study children in spring 2010. Assessment sections were similar to the fall sessions in that the order of the tests was varied. However, the spring battery was considerably shorter, taking on average 20–30 min. The fall testing sessions were longer than the spring sessions because more tests were given in the fall. In total, nine direct child assessments were given in the fall and four direct child assessments were given in the spring. Spring assessors also had to prove reliability on the battery of tests and show good rapport/child management skills in both simulated and real testing situations.

Also in spring 2010, independent assessors observed participating classrooms for 4–5 h. Care was taken that visits were representative of general classroom routines; visits were scheduled at times that were not disruptive to the teacher and on days that were typical of the usual environment for that classroom (i.e., not on a day when a field trip was planned, nor when the regular teacher was out sick). Assessors used the following three assessment tools: the Early Childhood Environment Rating Scale-Revised (ECERS-R; Harms, Clifford, & Cryer, 1998); the Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2008) and the Early Language and Literacy Classroom Observation (ELLCO; Smith, Dickinson, Sangeorge, & Anastasopoulos, 2002). One assessor completed all three assessments on the same classroom on the same day. The observation period began when the children arrived and continued until they left for the day. To ensure reliability, each assessor had to demonstrate at least 85% exact agreement with a master coder on each measure during one observation period. Assessors required between two and five observations to meet this criteria. Percent of exact agreement for the ECERS ranged from 85% to 93%, the percent of exact agreement for the CLASS ranged from 85% to 95% and the percent of exact agreement for the ELLCO ranged from 85% to 95%.

2.3. Child-level measures

2.3.1. Working memory

In both the fall and the spring, assessors administered two tests of children's working memory – Forward Digit Span (FDS; Gathercole & Pickering, 2000) and Backward Digit Span (BDS; Gathercole & Pickering, 2000). BDS and FDS measure different dimensions of working memory. BDS is considered a measure of the central executive component, while FDS is considered a measure of the phonological loop.

The tests are similar in structure. Before trial items are administered, the child has to pass a practice trial, demonstrating that he/she understands the directions of the task. If the child demonstrates adequate understanding, the assessor reads aloud a string of numbers to the test child, with approximately a 1-s pause between digits. The child then either has to repeat back exactly what the assessor said (in FDS) or reverse the string of numbers (in BDS). FDS is scored from one to six, while BDS is scored from one to five. The score represents the child's digit span memory (i.e. a two represents a digit span memory of two digits).

2.3.2. Cognitive inhibitory control

The Pencil Tapping task (PT; Diamond & Taylor, 1996) measures children's cognitive inhibitory control, defined as their ability to suppress prepotent responses. Children exhibiting cognitive inhibitory control can stop automatic responses and behaviors in favor of more appropriate ones. In the classroom, they can block distractions from peers and focus on learning (Carlson & Moses, 2001; Diamond, Carlson, & Beck, 2005).

In the Pencil Tapping task, children were asked to tap twice if the evaluator tapped once and tap once if the evaluator tapped twice. Assessors first administered a set of practice trials to ensure that children understood the rules of the task. Children who passed the practice were then given 16 total trials. To a lesser degree, this task also measures working memory and fine motor activity (Bierman et al., 2008). To reduce demands on fine motor skills, we substituted larger plastic kitchen spoons for pencils. Scores represented the correct number of trials out of 16.

2.3.3. Receptive vocabulary

Children's receptive vocabulary was measured using the Peabody Picture Vocabulary Test-III (PPVT-III; Dunn & Dunn, 1997), a nationally normed measure that has been widely used in diverse samples of young children (Love et al., 2005; Wong et al., 2008). The test has excellent split-half and test–retest reliability, as well as strong qualitative and quantitative validity properties. It requires children to choose which of four pictures best represents a stimulus word. In our analysis, we used standardized scores from the PPVT-III, where 100 represents the average age-adjusted child score. See Table 1 for summary statistics on the PPVT and other child-level outcome measures.

2.3.4. Child-level control variables

From district administrative records, we obtained information on children's race/ethnicity, home language, free/reduced lunch status, gender, and special-needs status. We used a vector of dichotomous indicators (*ASIAN, BLACK, HISPANIC, OTHER*, and *WHITE*, with *WHITE* as the reference group) to represent child race/ethnicity, each coded one when the child was from the specified group, zero otherwise. We also used a vector of dichotomous indicators (*ENGLISH, SPANISH, OTHER_LANGUAGE*, with *ENGLISH* as the reference group) to represent children's home language, each coded one when the child spoke the requisite language, zero otherwise. Because the district is divided into three attendance zones and because the demographic characteristics of students differ across these zones (McArdle, Osypuk, & Acevedo-Garcia, 2010), we used a vector of dichotomous indicators (*EAST, NORTH, WEST*, with *EAST* as the reference group) to represent children's district-defined attendance zone location, each coded one when the child lived within the specified attendance zone location, zero otherwise. To represent child free/reduced lunch status (*LUNCH*), gender (*MALE*), and special needs status (*SN*), we constructed dichotomous indicators, each coded one if the child falls into a certain demographic and zero otherwise. These covariates have been shown to predict children's early cognitive and educational outcomes in other studies, and there is a consensus in the early childhood education literature that these effects be controlled (Clements, Sarama, Spitler, Lange, & Wolfe, 2011; Wong et al., 2008). Using parent-reported data, we also controlled for children's previous care experience in the year prior to prekindergarten, using a vector of dichotomous indicators (Head Start, public centers, private centers, non-relative

Table 1
Descriptive statistics on child outcomes at the beginning of prekindergarten (fall) and end of prekindergarten (spring).

Construct	Assessment	Fall					Spring				
		N	Mean	*SD*	Min	Max	*N*	Mean	*SD*	Min	Max
Working memory	FDS	387	3.87	1.3	1	6	413	4.51	1.25	1	6
Working memory	BDS	365	1.27	0.58	1	3	412	1.44	0.72	1	4
Inhibitory control	PT	376	8.2	6.51	0	16	412	11.65	5.35	0	16
Receptive vocabulary	PPVT	391	88.19	17.63	40	136	411	94.45	17.89	11	139

home-based care, and relative care, with private centers as the reference group).

2.4. Classroom-level measures

2.4.1. Classroom quality

The Early Childhood Environment Rating Scale–Revised Edition (ECERS-R; Harms et al., 1998) is widely used in the early education field. The ECERS-R is organized into seven scales: space and furnishings, personal care routines, language reasoning, activities, interaction, program structure and parents and staff. A score of seven is considered "excellent," five is "good," three is "minimal," and one is "inadequate." Recent analysis of the ECERS-R found little support for the seven subscales, suggesting that the ECERS-R measures one, two or three quality factors (Cassidy et al., 2005; Gordon et al., 2010; Phillipsen, Burchinal, Howes, & Cryer, 1997). Accordingly, we fit a confirmatory factor analysis model testing the two factors that emerged in Pianta et al. (2005)'s study of 238 prekindergarten classrooms in six states. The two factors (and associated items) were: (1) Teaching and Interactions (encouraging children to communicate, using language to develop reasoning skills, general supervision of children, discipline, and staff–child interactions) and (2) Provisions for Learning (furnishings, room arrangement, gross motor equipment, art, blocks, dramatic play, and nature or science). Fit of the model was adequate ($\chi^2 = 81.541$, $p < .01$, $CFI = .90$; $TLI = 0.88$; $RMSEA = .08$; $SRMR = .08$). In analyses, we used the unit-weighted average of the items in each scale. Cronbach's alpha for the Teaching and Interactions construct was 0.79 and for the Provisions for Learning construct, 0.65. As a robustness check, given the noted variability in findings regarding the psychometric validity of the ECERS-R, we also used the ECERS total score, calculated in accordance with developer guidelines. The correlation between the Interactions and Provisions scores was 0.43 ($p < .001$), between Interactions and ECERS total, 0.79 ($p < .001$), and between Provisions and ECERS total, 0.70 ($p < .001$).

The Classroom Assessment Scoring System (CLASS; Pianta et al., 2008) focuses on interactions between students and teachers. Like the ECERS-R, the CLASS is measured on a seven-point scale with higher scores indicating higher quality. A score of seven is considered "excellent," five is "good," three is "minimal," and one is "inadequate." The CLASS consists of three subscales – Emotional Climate, Instructional Support, and Classroom Organization – and each subscale is composed of multiple indicators. In our study, observations for the CLASS were conducted in 30-min cycles for observation and scoring. As there is some consensus in the field regarding the psychometric structure of the CLASS (Hamre et al., submitted for publication; Pakarinen et al., 2010), we assessed the psychometrics of the CLASS using CFA models within our prekindergarten sample. A three-factor model showed adequate fit to the data ($\chi^2 = 57.73$, $p < .01$, $CFI = 0.97$, $TLI = 0.94$, $RMSEA = 0.11$, $SRMR = 0.06$) and showed statistically significantly superior fit to the data than a two-factor solution (Instructional Support and Emotional Climate/Classroom Organization; χ^2 difference = 1.36, df = 1). In our regression models, we thus used the three CLASS scales as described in the CLASS manual.

The Early Language and Literacy Classroom Observation tool (ELLCO; Smith et al., 2002) is used to evaluate the environment and teaching practices of classrooms in regards to language and literacy. It is comprised of three scales: a literacy environment checklist, a general classroom environment quality scale, and a language, literacy, and curriculum quality scale. We used only the eight-item literacy activities rating scale, as in Burchinal, Xue, Tien, Auger, and Mashburn (2011), as the literacy environment checklist does not measure quality and as the classroom observation had considerable conceptual overlap with the ECERS-R. Ratings on the literacy activities scale are determined as follows: a score of five is considered "exemplary," three is "basic," and one is "deficient." Smith et al. (2002) report that the Cronbach's alpha for the literacy activities scale is 0.86. We conducted CFA analysis on the eight literacy activities rating scale items and found good fit ($\chi^2 = 30.14$, $CFI = 0.96$, $TLI = 0.94$, $RMSEA = 0.08$, $SRMR = 0.05$). Cronbach's alpha was 0.85. In all analyses, we used a unit-weight average of the eight literacy scale items (*ELLCO Literacy*). For descriptive statistics on the ECERS, CLASS and ELLCO predictors, see Table 2.

2.4.2. Classroom-level control variables

We controlled for several classroom-level variables believed to be important in early childhood classrooms. In accordance with the National, Institute for Early Education Research (Barnett et al., 2010) recommendations, we included a dichotomous variable indicating whether class size was less than 20 or not, even though empirical evidence on the importance of class size is mixed (Blau, 1999; Mashburn et al., 2008; Zaslow et al., 2011). In the BPS, the maximum prekindergarten class size is 22.

Using district administrative records, we also created a dichotomous variable set equal to one if the teacher had a masters degree and zero otherwise. Some previous studies have found teacher education to be uncorrelated with child outcomes in preschool (Early et al., 2007; Mashburn et al., 2008). However, we include this variable for several reasons: (1) the National, Institute for Early Education Research recommends all preschool teachers hold at least a B.A (Barnett et al., 2010) and (2) the BPS is somewhat unique in this literature in that all prekindergarten teachers hold a bachelor's degree, the majority hold masters degrees, and all are subject to the same requirements as K-12 teachers.

Table 2
Descriptive statistics (means and *SD*s) for the quality indicators ($N = 83$).

	Mean	*SD*	Range (Min–Max)
Quality of the overall environment			
ECERS Interactions	5.54	1.21	2.20–7.00
ECERS Provisions	3.72	0.55	1.86–4.86
ECERS total	4.47	0.50	3.29–5.60
Quality of teacher-child interactions			
CLASS Emotional Support	5.63	0.60	4.00–6.83
CLASS Instructional support	4.30	0.84	2.22–5.67
CLASS Classroom organization	5.10	0.68	2.75–6.22
Quality of language and literacy			
ELLCO Language, Literacy, and Curriculum	3.53	0.45	2.50–4.50

Note: All constructs are rated on a 1–7 scale, except ELLCO Language, Literacy, and Curriculum which is rated on 1–5 scale.

2.5. Data analytic approach

To examine the relationship between continuous classroom quality measures and child outcomes, we fit a series of regression models. We first modeled the relationship between each of the six continuous quality indicators from the CLASS, ECERS-R, and ELLCO and each child outcome score as linear, controlling for child pretest scores, child demographic characteristics, teacher education, and classroom size. In a second model and in accordance with recent findings that suggest that such relationships may be curvilinear (Burchinal, Kainz, et al., 2011), we included a quadratic term for each continuous quality indicator. We also fit spline regression models, in which we estimated two linear slopes – one slope for the lower-quality range and one slope for the higher-quality range. We determined the "knot," or threshold at which the slope was allowed to change in two ways. First, for statistically significant quadratic relationships, we calculated the inflection point of the curve. That is, we applied basic Calculus to our results from our statistically significant quadratic models to determine the point at which the slope changed in sign and we chose that point as the cutpoint in our threshold analysis. As a second strategy, we began with conceptually chosen cutpoints that have been used elsewhere in this literature (5 for CLASS Emotional Support and Classroom Organization, 2.75 for CLASS Instructional Support, 4.5 for the ECERS scales, and 4 for the ELLCO Literacy scale; see Burchinal et al., 2010; Burchinal, Xue, et al., 2011). However, because of the distribution of our data, we chose different cutpoints for some of our quality indicators. In our data, only 12% of classrooms ($N = 10$) had Emotional Support scores below 5. Accordingly, we chose a cutpoint of 5.5, as 33% of classrooms ($N = 27$) averaged below 5.5. Likewise, as noted earlier in this paper, BPS Instructional Support scores are much higher than average. Only 5% of BPS classrooms scored below 2.75 on this variable, so we chose a cutpoint of 5, which conceptually represents the "Good" range on the CLASS scale. Overall, 80% of classrooms ($N = 66$) scored below a 5 on Instructional Support. For similar distributional reasons, for the ECERS, we chose a cutpoint of 4 for ECERS Provisions for Learning ($N = 46$ or 55% classrooms below this threshold) and for ECERS Interactions and Teaching, we chose a cut point of 5 ($N = 21$ or 25% below this threshold). Number and percentage of classrooms falling below conceptually chosen cutpoints that have been used elsewhere in this literature were: ECERS total cutpoint of 4.5, $N = 41$, 49% of classrooms; ELLCO cutpoint of 4, $N = 66$, 80% of classrooms; and CLASS Classroom Organization cutpoint of 5, $N = 27$, 33% of classrooms.

Overall, chosen cut points correspond approximately to the "Good" threshold for all three quality measures. There is no agreed-upon best method for choosing the knot and possibilities include using the inflection point, conceptually defined points, empirically identified points, using non-linear regression methods, and/or some combination of these approaches (Marsh & Cormier, 2002; Zaslow et al., 2010). We choose conceptually defined points and inflection points because our spline models were intended only to help us interpret the magnitude of the detected curvilinear relationships in higher- and lower-quality ranges and to compare to previous literature regarding whether relationships are stronger in higher- versus lower-quality classrooms (Burchinal et al., 2010; Burchinal, Xue, et al., 2011). Although our results do provide some limited evidence regarding how results are sensitive to the method by which spline knots are chosen, comparison of methods for choosing the spline knot is outside of the scope of the present paper.

All regression analysis for both research questions was conducted in Stata 11 and we used multiple imputation to account for missing covariates, fall test, and spring test data. We assumed that data are missing at random, and we were guided in our implementation of multiple imputation by Graham (2009) who argues that, even if the MAR assumption is violated, "MI and ML methods are always at least as good as the old procedures. . . and MI/ML are typically better than old methods, and often very much better" (p. 559). We adjusted for the nesting of students within schools and classrooms using robust standard errors.

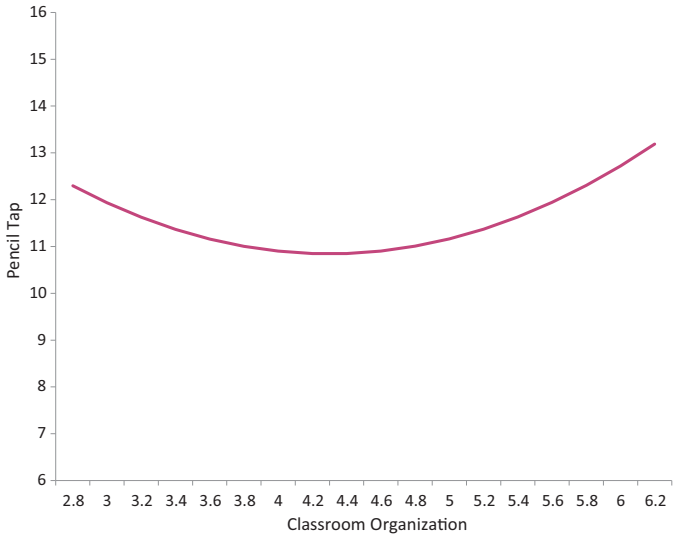

Fig. 1. Plot of the fitted relationship between child Pencil Tap scores and CLASS Classroom Organization. *Note*: Fitted values obtained by substituting in average values for each control variable. Values of CLASS Classroom Organization were chosen based on the distribution of the scale in the sample.

3. Results

3.1. Descriptive statistics – correlations between classroom quality predictors

As shown in Table 3, CLASS, ECERS and ELLCO quality rating scales were moderately to highly and statistically significantly correlated, ranging from 0.39 (ECERS Provisions and CLASS Instructional Support) to 0.86 (CLASS Emotional Support and CLASS Classroom Organization). Scales generally were more highly correlated within the same measure versus across different measures.

3.2. Linear and quadratic model results

In Table 4, we display results from modeling the relationship between quality indicators and children's receptive vocabulary and executive function outcomes as both linear and quadratic, controlling for covariates and the relevant fall pretest, adjusting for the clustering of students in schools, and using multiple imputation for missing data. We translated detected statistically significant linear relationships into effect sizes (Cohen's d) by multiplying the predictor's coefficient by the standard deviation of the predictor and dividing by the standard deviation of the outcome (Burchinal et al., 2010; NICHD & Duncan, 2003; Rimm-Kaufman et al., 2009).

As shown in Table 4, none of the quality predictors were statistically significantly associated with child PPVT-III standardized scores. Also shown in Table 4, the ELLCO Literacy scale showed a positive, linear association with children's Pencil Tap scores ($\beta = 1.05$; $p < .05$). This translated into a small effect size of 0.07. There were no statistically significant linear relationships between the Pencil Tap and the CLASS scales, but the three CLASS predictors each showed quadratic, statistically significant associations with children's Pencil Tap scores ($p < .05$). For parsimony, we included a plot of the fitted relationship between CLASS predictors and Pencil Tap scores for Classroom Organization only (see Fig. 1). As shown in Fig. 1, the relationship between Classroom Organization

Table 3
Correlations between classroom quality predictors ($N = 83$).

	1	2	3	4	5	6
1. CLASS Instructional support	1.00					
2. CLASS Emotional support	0.69***	1.00				
3. CLASS Classroom organization	0.71***	0.86***	1.00			
4. ECERS Interactions	0.60***	0.67***	0.59***	1.00		
5. ECERS Provisions	0.39***	0.59***	0.47***	0.43***	1.00	
6. ECERS total Overall structural quality	0.50***	0.65***	0.52***	0.79***	0.70***	1.00
7. ELLCO Language, literacy, and curriculum	0.65***	0.69***	0.70***	0.58***	0.61***	0.64

*** $p < .001$.

and Pencil Tap scores appears to be stronger at higher levels of Classroom Organization. Plots of the relationships between Pencil Tap scores and the other two CLASS predictors were similar to Fig. 1, with relationships between those scales and Pencil Tap scores also stronger at higher levels of those quality scales (results available upon request). To check the sensitivity of detected quadratic relationships to outliers, we refit Pencil Tap and CLASS models, dropping cases in which the observed quality rating score was ±2 *SD*s from the mean for that quality rating score (Classroom Organization, $N = 16$ children in 4 classrooms; Instructional Support, $N = 19$ children in 4 classrooms; Emotional Support, $N = 24$ children in five classrooms). Results for the Pencil Tap and Emotional Support and the Pencil Tap and Instructional Support were very similar in magnitude, direction, and statistical significance to those shown in Table 4. Results for the Pencil Tap and Classroom Organization were similar in magnitude and direction to those in Table 4 but the quadratic term was no longer statistically significant ($p = .06$).

Table 4 also displays results for the two working memory outcomes – Forward Digit Span and Backward Digit Span. There were no statistically significant associations (either linear or quadratic) between quality predictors and the working memory outcomes.

3.3. Spline regression model results

In Table 5, we display the results from applying spline regression methods, with the spline knot set equal to inflection-point-determined cutpoints and conceptually determined cut points. The second column of the table corresponds to slopes and standard errors for classrooms in the lower-quality range, while the third column corresponds to slopes and standard errors for classrooms in the higher-quality range. The fourth column indicates whether the difference between the slopes for the lower- and higher-quality classrooms was statistically significantly different ($p < .05$). Inflection-point-determined spline knots were set as follows for the Pencil Tap: Instructional Support = 3.90; Emotional Support = 5.13; Classroom Organization = 4.29. The percentage of classrooms with ratings falling below the spline knot value were as follows: Instructional Support, 28% ($N = 23$); Emotional Support, 14% ($N = 12$); Classroom Organization, 13% ($N = 11$). (For information on spline knot locations and percentage of classrooms below these values for conceptually determined cutpoints, see the "Data Analytic Approach" section of this paper.)

Table 4
Results of regressing receptive vocabulary and executive function measures on classroom quality indicators.

	PPVT-III		Pencil Tap		Forward Digit Span		Backward Digit Span	
	Linear	Quad.	Linear	Quad.	Linear	Quad.	Linear	Quad.
CLASS Instructional Support	0.13	7.01	0.16	−4.83*	−0.03	−0.09	−0.04	0.33
	(0.93)	(6.50)	(0.30)	(1.81)	(0.06)	(0.54)	(0.04)	(0.32)
CLASS Instructional Support^2	–	−0.85	–	0.62*	–	0.01	–	−0.05
	–	(0.83)	–	(0.23)	–	(0.07)	–	(0.04)
CLASS Emotional Support	−1.23	−24.58	0.56	−9.03*	−0.09	−1.01	−0.04	−0.06
	(1.14)	(13.51)	(0.39)	(4.34)	(0.01)	(1.18)	(0.05)	(0.75)
CLASS Emotional Support^2	–	2.15	–	0.88*	–	0.08	–	0.00
	–	(1.29)	–	(0.40)	–	(0.11)	–	(0.07)
CLASS Organization	−0.87	−3.76	0.66	−5.57∼	0.00	0.14	−0.02	0.276
	(0.92)	(9.01)	(0.36)	(2.82)	(0.08)	(0.67)	(0.04)	(0.40)
CLASS Organization^2	–	0.30	–	0.65*	–	−0.01	–	−0.03
	–	(0.99)	–	(0.31)	–	(0.07)	–	(0.04)
ECERS Provisions	−0.41	−14.70	0.49	−1.79	−0.19	−0.66	−0.02	−0.02
	(1.20)	(8.50)	(0.34)	(2.65)	(0.10)	(0.69)	(0.05	(0.44)
ECERS Provisions^2	–	2.09	–	0.33	–	0.07	–	0.00
	–	(1.31)	–	(0.40)	–	(0.10)	–	(0.07)
ECERS Interactions	−0.12	−4.49	0.09	−0.64	−0.02	−0.29	−0.03	0.13
	(0.53)	(3.74)	(0.19)	(1.30)	(0.04)	(0.35)	(0.03)	(0.21)
ECERS Interactions^2	–	0.43	–	0.07	–	0.03	–	−0.02
	–	(0.36)	–	(0.13)	–	(0.03)	–	(0.02)
ECERS total	1.77	−15.17	0.37	−6.96	−0.12	0.80	−0.06	−0.03
	(1.35)	(19.54)	(0.47)	(5.43)	(0.12)	(1.42)	(0.06)	(1.02)
ECERS total^2	–	1.93	–	0.84	–	−0.11	–	−0.00
	–	(2.24)	–	(0.62)	–	(0.16)	–	(0.12)
ELLCO Literacy	0.32	−25.80	1.05*	−1.98	−0.08	−0.11	−0.11	0.99
	(1.54)	(17.81)	(0.51)	(7.28)	(0.15)	(1.98)	(0.06)	(0.93)
ELLCO Literacy^2	–	3.73	–	0.43	–	0.00	–	−0.16
	–	(2.58)	–	(1.04)	–	(0.28)	–	(0.13)

Note: All models control for child demographic characteristics (race/ethnicity, gender, home language, free/reduced lunch, special needs, age, attendance zone, and pre-BPS childcare experience), child pretest scores, teacher masters degrees and class size and include multiple imputation for missing data. Robust standard errors were used to adjust for clustering at the school level.

* $p < .05$.

Table 5
Spline regression slopes (and standard errors), using the inflection-point and conceptually defined spline knots.

	Low quality classrooms	High quality classrooms	Slopes statistically significantly different?
Inflection-point determined spline knots			
Pencil Tap			
CLASS Instructional Support	−1.30**	1.21*	**
	(0.47)	(0.48)	
CLASS Emotional Support	−1.00	1.20*	
	(0.90)	(0.58)	
CLASS Classroom Organization	−0.92	1.09*	
	(0.92)	(0.50)	
Conceptually defined spline knots			
PPVT-III			
CLASS Emotional Support	−3.80**	5.08	*
	(1.47)	(3.21)	

Note: All models control for child demographic characteristics (race/ethnicity, gender, home language, free/reduced lunch, special needs, age, attendance zone, and pre-BPS childcare experience), child pretest scores, teacher masters degrees and class size and include multiple imputation for missing data. Robust standard errors were used to adjust for clustering at the school level.

* $p < .05$.

** $p < .01$.

Among the inflection-point determined spline knot models, CLASS Instructional Support was negatively and statistically significantly related to Pencil Tap scores for children in lower- quality classrooms ($\beta = -1.30$; $p < .01$; $d = -0.20$) but positively and statistically significantly related to Pencil Tap scores for children in higher-quality classrooms ($\beta = 1.21$; $p < .05$; $d = 0.19$). CLASS Emotional Support and CLASS Organization both were predictive of Pencil Tap scores only among children in higher-quality classrooms (Emotional Support: $\beta = 1.20$; $p < .05$; $d = 0.14$; Organization: $\beta = 1.09$; $p < .05$; $d = 0.14$). We fit models with conceptually defined cutpoints for each child outcome and quality measure combination (4 outcomes and 7 quality measures = 28 combinations in total). For parsimony, in Table 5, we display only those for which there was evidence of a stronger or weaker relationship in the higher or lower ranges of quality. The only child outcome and quality measure combination to show evidence of a threshold effect was the PPVT-III and CLASS Emotional Support. CLASS Emotional Support was negatively associated with PPVT-III scores only among children in lower-quality classrooms ($\beta = -3.80$; $p < .05$; $d = -0.13$).

4. Discussion

Taken together, our results are largely consistent with previous findings in the literature that suggest that classroom quality indicators have small or null associations with gains in children's developmental outcomes in preschool (Burchinal et al., 2010; Burchinal, Kainz, et al., 2011; NICHD & Duncan, 2003; Zaslow et al., 2010, 2011). In linear models, we found null associations between changes in children's receptive vocabulary and seven quality rating scales from three different classroom quality measures. Findings from linear models for executive function and the quality ratings scales were similarly weak or null. We detected several statistically significant quadratic relationships between gains on the Pencil Tap and each of the three CLASS scales, with relationships stronger at higher levels of those quality scales. In spline regression models, we found that the relationship between gains on the Pencil Tap and the three CLASS scales was of greater magnitude in higher-quality versus lower-quality classrooms, while the opposite was true for receptive vocabulary and the CLASS Emotional Support scale. The latter result differs somewhat from Burchinal et al. (2010) who found no threshold effect for receptive vocabulary and CLASS Emotional Support. Consistent with Burchinal et al. (2010), effect sizes were all in the small range for our spline regression models; we found no evidence of a threshold of quality above or below which child outcomes were particularly strongly associated with classroom quality.

Contrary to one hypothesis in the literature on older students (Rimm-Kaufman et al., 2009), we did not find that classroom organization was a stronger predictor of children's executive function (inhibitory control or working memory), compared to other classroom quality measures. Further, we found that CLASS scores were more strongly associated with inhibitory control in higher-quality classrooms, compared to lower-quality classrooms. We hypothesize that classrooms with higher process quality, as measured by the CLASS, may provide children with more opportunities to practice inhibitory control. This hypothesis is somewhat in line with the mechanism proposed by Bodrova and Leong (2006) – that higher-quality classrooms may build children's EF by supporting child activity choice and reflection. It should also be noted that the Boston program implemented consistent mathematics curricula across prekindergarten classrooms. Prekindergarten mathematics tasks typically require children to inhibit automatic or prepotent responses (Blair, Gamson, Thorne, & Baker, 2005); thus a possible explanation in this context is that classrooms with higher process quality may have spent more time on mathematics activities.

Our finding that working memory showed no statistically significant relationship to any quality measure was somewhat surprising, given that classrooms with higher instructional quality would likely make more demands on children's working memory. This lack of association too should be considered in light of the mathematics curriculum in place in study classrooms. To be able to correctly solve prekindergarten mathematics activities, a child must hold multiple heuristics or strategies in memory (Blair et al., 2005), leading some to hypothesize that prekindergarten mathematics activities may build children's working memory (Welsh et al., 2010). Future work is needed at the preschool level to examine both the possible mechanisms by which classroom quality may affect children's executive functions, the ways in which mechanisms may differ by quality dimension, and whether the associations detected in our study are atypical or consistent with those in other contexts.

Interestingly, in our study and contrary to one hypothesis in the literature and to some empirical findings (Burchinal, Kainz, et al., 2011), better alignment between the child developmental domain and the quality rating scale did not seem to matter. That is, the conceptual fit of the examined outcome and the rating scale did not result in stronger detected relationships. For example, we found that the ELLCO Literacy scale, which purports to measure the instructional quality of language and literacy instruction, was not predictive of children's receptive vocabulary outcomes in our study, while it did predict children's cognitive inhibitory control

(effect size = 0.07). Similarly, in conceptually chosen spline regression models, we found no relationship between higher or lower values of CLASS Instructional Support and PPVT-III scores, while we did find such evidence for the CLASS Emotional Support and PPVT-III scores, even though theory and other empirical research suggest that the former are conceptually more aligned than the latter. We note, however, that the CLASS Emotional Support scale includes items that focus on communications between teachers and students. Higher Emotional Support scores thus may be linked to a richer language environment.

We found limited support for another hypothesis in the literature – that weak relationships between quality rating scales and child outcomes might be explained by range restriction, particularly for the CLASS Instructional Support scale. While we did not observe ceiling effects for the Instructional Support scale (the maximum score was 5.67 out of a possible 7 in our sample) and it is possible that scores in the 6–7 range are required to detect moderate-to-large associations, the Boston Public Schools program we studied is of higher average instructional quality than many other prekindergarten programs. Boston rated higher on the CLASS Instructional Support scale than a national sample of 671 publicly funded prekindergarten programs (Mashburn et al., 2008). Boston also rated higher than another public prekindergarten program in Tulsa that has shown strong, positive impacts on children (Gormley et al., 2005); CLASS Instructional Support scores in Boston averaged 4.30 while the Tulsa average was 3.21 (Phillips, Gormley, & Lowenstein, 2009). Our results do not rule out the possibility of range restriction entirely, and our sample is fairly small. Another, larger dataset with classrooms that span the full range of values for the CLASS Instructional Support scale would better address the range restriction hypothesis. However, finding a program with Instructional Support scores higher than Boston, with its majority masters-level teachers and its mathematics and literacy curricula, is no easy task in practical terms.

Most compelling to us in this literature, although our data do not speak directly to it, is the hypothesis put forth by Burchinal et al. (2010) and Zaslow et al. (2010) that the current measures of quality may simply not be strong measures of the classroom quality factors that improve children's academic outcomes. The psychometric properties of the ECERS-R in particular raise questions (Cassidy et al., 2005; Gordon et al., 2010) and while the psychometric evidence for the CLASS and the ELLCO Literacy scale is stronger, their evidence base too is still emerging. For example, using IRT methods, Gordon et al. (2010) identified ways in which ECERS-R items are disordered, but to our knowledge, such research has yet to be replicated on the CLASS or the ELLCO Literacy scale. Future such sophisticated work on the psychometrics of commonly used quality measures is needed, particularly given that these measures increasingly are being used at the policy level in initiatives like Quality Rating Improvement Systems to evaluate programs and to allocate public resources (Tout, Zaslow, Halle, Forry, & Child, 2009).

In addition, how quality data were collected in our study may have impacted our results. Observers in our study completed all three quality measures in the same 4–5 h visit. One-time measurement of a construct as complex and multi-faceted as classroom quality may not be adequate. Adding to potential measurement error problems, teachers may also have changed their behavior during the observation. More frequent measurement of classroom quality using these observational measures may be needed to obtain more reliable measures of classroom quality.

Finally, it should be noted that although observational quality measures showed null or weak relationships with children's outcomes in our study, the BPS has found these measures critical for teacher professional development and facility improvement efforts. Beginning in 2006, the BPS shared scores on these measures with prekindergarten teachers, and as part of their work with teachers, coaches addressed areas in which their scores showed room for improvement. Average scores on the ECERS-R and CLASS improved over time and overall, a large causal evaluation found that the BPS prekindergarten substantially improves children's school readiness and that some impacts are larger than any other similarly evaluated prekindergarten program (Weiland & Yoshikawa, in press). While there were other quality supports put into place at the same time that may also likely have contributed to the increase in quality (such as the chosen curricula), it could be the case thus that such observational quality tools are valid and useful for certain purposes (like guiding teacher professional development and improving teacher practice), even if they are not strongly associated with gains in children's outcomes during the prekindergarten year. It is also possible that attaining a certain level of quality is a necessary but not sufficient condition for obtaining impacts on children's cognitive development.

4.1. Limitations of the present study

Our study has several important limitations. First, the work presented is exploratory and non-causal in nature. While we controlled for student, teacher, and classroom characteristics, it is possible that our results are spurious, due to selection bias or omitted variables bias. Second, our child and classroom sample is relatively small, compared to other similar studies examining the relationship between child outcomes and quality indicators. We thus may have failed to detect relationships that do exist in the data due to statistical power issues. Third, our study setting is an unusual one in the preschool and prekindergarten context. Few programs have such a high percentage of masters-level teachers or a consistent mathematics and literacy curricula implemented with moderate-to-good fidelity in place across a large-scale program. Also, all programs are based in the same city, entirely within urban public schools, and not across different program auspices, as in many other prekindergarten programs. As such, the external validity of our results may be limited. Fourth, children in our study differed from non-participating children from the same district on 8 out of 20 examined background characteristics. While most of these differences represented a difference of a few percentage points, results reported here may not generalize to the full study district. Fifth, the fall child testing session was longer than the spring session, meaning that our fall outcome measures may be subject to more measurement error than our spring measures due to child fatigue. Finally, in accordance with recent advice in the literature (Schochet, 2009), we did not adjust the statistical significance of reported results to account for Type I error, as our study was exploratory in nature.

References

Ashe, M. K., Reed, S., Dickinson, D. K., Morse, A. B., & Wilson, S. J. (2009). Opening the World of Learning: Features, effectiveness, and implementation strategies. *Early Childhood Services*, *3*, 179–191.

Auger, A., Farkas, G., Duncan, G. J., Burchinal, M., & Vandell, D. L. (2012, November). *Child care quality and academic achievement: Results from PCER*. Paper presented at the annual meeting of the Association for Public Policy Analysis and Management, Baltimore, MD.

Barnett, W., Jung, K., Yarosz, D. J., Thomas, J., Hornbeck, A., Stechuk, R., et al. (2008). Educational effects of the Tools of the Mind curriculum: A randomized trial. *Early Childhood Research Quarterly*, *23*(3), 299–313.

Barnett, W. S., Epstein, D. J., Carolan, M. E., Fitzgerald, J., Ackerman, D. J., & Friedman, A. H. (2010). *The state of preschool 2010*. Retrieved from the National, Institute for Early Education Research website http://nieer.org/sites/nieer/files/yearbook.pdf

Bierman, K. L., Domitrovich, C. E., Nix, R. L., Gest, S. D., Welsh, J. A., Greenberg, M. T., et al. (2008). Promoting academic and social–emotional school readiness: The Head Start REDI program. *Child Development*, *79*(6), 1802–1817.

Blair, C. (2002). School readiness: Integrating cognition and emotion in a neurobiological conceptualization of children's functioning at school entry. *American Psychologist*, *57*(2), 111–127.

Blair, C., Gamson, D., Thorne, S., & Baker, D. (2005). Rising mean IQ: Cognitive demand of mathematics education for young children, population exposure to formal schooling, and the neurobiology of the prefrontal cortex. *Intelligence*, *33*(1), 93–106.

Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development*, *78*(2), 647–663.

Blau, D. M. (1999). The effect of child care characteristics on child development. *Journal of Human Resources*, *34*(4), 786–822.

Bodrova, E., & Leong, D. J. (2006). Self-regulation as a key to school readiness: How early childhood teachers can promote this critical competency. In M. Zaslow, & I. Martinez-Beck (Eds.), *Critical issues in early childhood professional development* (5th ed., pp. 203–224). Baltimore, MD: Brookes.

Bronfrenbrenner, U., & Morris, P. A. (1998). The ecology of developmental processes. In W. Damon, & R. M. Lerner (Eds.), *Handbook of child psychology: Theoretical models of human development* (pp. 993–1028). New York, NY: Wiley.

Burchinal, M., Howes, C., Pianta, R., Bryant, D., Early, D., Clifford, R., et al. (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.

Burchinal, M., Kainz, K., & Cai, Y. (2011). How well do our measures of quality predict child outcomes? A meta-analysis and coordinated analysis of data from large-scale studies of early childhood settings. In M. Zaslow, I. Martinez-Beck, K. Tout, & T. Halle (Eds.), *Quality measurement in early childhood settings* (pp. 11–31). Baltimore, MD: Brookes.

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.

Burchinal, M., Xue, Y., Tien, H., Auger, A., & Mashburn, A. (2011, April). *Testing for thresholds in associations between child care quality and child outcomes*. Paper presented at the Society for Research in Child Development, Montreal, Canada.

Campbell, F. A., & Ramey, C. T. (1994). Effects of early intervention on intellectual and academic achievement: A follow-up study of children from low-income families. *Child Development*, *65*(2), 684–698.

Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development*, *72*(4), 1032–1053.

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*(3), 345–360.

Clements, D. H., & Sarama, J. (2007a). *SRA real math, prek-building blocks*. Columbus, OH: SRA/McGraw-Hill.

Clements, D. H., & Sarama, J. (2007b). Effects of a preschool mathematics curriculum: Summative research on the Building Blocks project. *Journal for Research in Mathematics Education*, *38*(2), 136–163.

Clements, D. H., Sarama, J. H., Spitler, M. E., Lange, A. A., & Wolfe, C. B. (2011). Mathematics learned by young children in an intervention based on learning trajectories: A large-scale cluster randomized trial. *Journal for Research in Mathematics Education*, *42*(2), 127–166.

Diamond, A., Carlson, S. M., & Beck, D. M. (2005). Preschool children's performance in task switching on the dimensional change card sort task: Separating the dimensions aids the ability to switch. *Developmental Neuropsychology*, *28*(2), 689–729.

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), 315–344.

Dickinson, D. K., Kaiser, A., Roberts, M., Hofer, K. G., Darrow, C. L., & Griffenhagen, J. B. (2011, March). *The effects of two language focused preschool curricula on children's achievement through first grade*. Paper presented at the Spring Conference of the Society for Research in Educational Effectiveness, Washington, DC.

Dickinson, D. K., & Smith, M. W. (2001). Supporting language and literacy development in the preschool classroom. In D. K. Dickinson, & P. O. Tabors (Eds.), *Beginning literacy with language: Young children learning at home and school* (pp. 139–148). Baltimore, MD: Brookes.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody picture vocabulary test-third edition*. Bloomington, MN: Pearson Assessments.

Early, D. M., Maxwell, K. L., Burchinal, M., Bender, R. H., Ebanks, C., Henry, G. T., et al. (2007). Teachers' education, classroom quality, and young children's academic skills: Results from seven studies of preschool programs. *Child Development*, *78*(2), 558–580.

Fuhs, M. E., & Day, J. D. (2011). Verbal ability and executive functioning development in preschoolers at Head Start. *Developmental Psychology*, *47*(2), 404–416.

Gathercole, S. E., & Pickering, S. J. (2000). Working memory deficits in children with low achievements in the national curriculum at 7 years of age. *British Journal of Education Psychology*, *70*(2), 177–194.

Gordon, R. A., Fujimoto, K., Kaestner, R., Korenman, S., & Abner, K. (2010, June). *Psychometric properties of the ECERS-R in a national sample of four year olds*. Poster presented at the Institute of Education Sciences Research Conference, National Harbor, MD.

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.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*(1), 549–576.

Hamre, B. K., Pianta, R. C., Downer, J. T., Hakigami, A., Mashburn, A. J., Jones, S., et al. Teaching through interactions: Testing a developmental framework for understanding teacher effectiveness in over 4,000 U.S. early childhood and elementary classrooms, submitted for publication.

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

Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R., et al. (2008). Ready to learn? Children's pre-academic achievement in pre-kindergarten programs. *Early Childhood Research Quarterly*, *23*(1), 27–50.

Howse, R. B., Lange, G., Farran, D. C., & Boyles, C. D. (2003). Motivation and self-regulation as predictors of achievement in economically disadvantaged young children. *Journal of Experimental Education*, *71*(2), 151–174.

Hustedt, J. T., Barnett, W. S., Jung, K., & Goetze, L. D. (2009). *The New Mexico preK evaluation: Results from the initial four years of a new state preschool initiative – Final report*. Retrieved from the National, Institute for Early Education Research website http://nieer.org/pdf/new-mexico-initial-4-years.pdf

Hustedt, J. T., Barnett, W. S., Jung, K., & Thomas, J. (2007). *The effects of the Arkansas Better Chance Program on young children's school readiness*. Retrieved from the State of Arkansas website http://www.arkansas.gov/childcare/abc/pdf/longreport.pdf

Love, J., Ross, C., Raikes, H., Constantine, J., Boller, K., Brooks-Gunn, J., et al. (2005). The effectiveness of early Head Start for 3-year-old children and their parents: Lessons for policy and programs. *Developmental Psychology*, *41*(6), 885–901.

Marsh, L. C., & Cormier, D. R. (2002). *Spline regression models*. London, UK: Sage Publications.

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., et al. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development*, *79*(3), 732–749.

McArdle, N., Osypuk, T., & Acevedo-Garcia. (2010). Prospects for equity in Boston Public Schools' school assignment plans. Retrieved from http://diversitydata.sph.harvard.edu/Publications/Prospects_for_Equity_in%20_Boston_Schools.pdf

McClelland, M. M., Acock, A. C., & Morrison, F. J. (2006). The impact of kindergarten learning related skills on academic trajectories at the end of elementary school. *Early Childhood Research Quarterly*, *21*(4), 471–490.

McClelland, M. M., Cameron, C. E., Connor, C. M., Farris, C. L., Jewkes, A. M., & Morrison, F. J. (2007). Links between behavioral regulation and preschoolers' literacy, vocabulary, and math skills. *Developmental Psychology*, *43*(4), 947–959.

McClelland, M. M., Morrison, F. J., & Holmes, D. L. (2000). Children at risk for early academic problems: The role of learning-related social skills. *Early Childhood Research Quarterly*, *15*(3), 307–329.

National Institute of Child Health and Human Development Early Child Care Research Network (NICHD ECCRN). (2006). Child-care effect sizes for the NICHD Study of Early Child Care and Youth Development. *American Psychologist*, *61*(2), 99–116.

National Institute of Child Health and Human Development Early Child Care Research Network (NICHD ECCRN), & Duncan, G. J. (2003). Modeling the impacts of child care quality on children's preschool cognitive development. *Child Development*, *74*(5), 1454–1475.

Neuman, S. B., & Carta, J. J. (2011). Advancing the measurement of quality for early childhood programs that support early language and literacy development. In M. Zaslow, I. Martinez-Beck, K. Tout, & T. Halle (Eds.), *Quality measurement in early childhood settings* (pp. 51–76). Baltimore, MD: Brookes.

Pakarinen, E., Lerkkanen, M., Poikkeus, A., Kiuru, N., Siekkinen, M., Rasku-Puttonen, H., et al. (2010). A validation of the classroom assessment scoring system in Finnish kindergartens. *Early Education and Development*, *21*(1), 95–124.

Peisner-Feinberg, E. S., Burchinal, M. R., Clifford, R. M., Culkin, M. L., Howes, C., Kagan, S. L., et al. (2001). The relation of preschool child-care quality to children's cognitive and social development trajectories through second grade. *Child Development*, *72*, 1534–1553.

Phillips, D., Gormley, W. T., & Lowenstein, A. (2009). Inside the pre-kindergarten door: Classroom climate and instructional time allocation in Tulsa's pre-K programs. *Early Childhood Research Quarterly*, *24*(3), 213–228.

Phillipsen, L. C., Burchinal, M. R., Howes, C., & Cryer, D. (1997). The prediction of process quality from structural features of child care. *Early Childhood Research Quarterly*, *12*, 281–303.

Pianta, R., Howes, C., Burchinal, M., Bryant, D., Clifford, R., Early, D., et al. (2005). Features of pre-kindergarten programs classrooms, and teaches: Do they predict observed classroom quality and child–teacher interactions? *Applied Developmental Science*, *9*(3), 144–159.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system [CLASS] manual: Pre-K*. Baltimore, MD: Brookes.

Rimm-Kaufman, S. E., Curby, T. W., Grimm, K. J., Nathanson, L., & Brock, L. L. (2009). The contribution of children's self-regulation and classroom quality to children's adaptive behaviors in the kindergarten classroom. *Developmental Psychology*, *45*(4), 958–972.

Schickedanz, J., & Dickinson, D. (2005). *Opening the world of learning*. Iowa City, IA: Pearson Publishing.

Schochet, P. Z. (2009). An approach for addressing the multiple testing problem in social policy impact evaluations. *Evaluation Review*, *33*, 539–567.

Schweinhart, L. J., Montie, J., Xiang, Z., Barnett, W. S., Belfield, C. R., & Nores, M. (2005). *Lifetime effects: The High/Scope Perry Preschool study through age 40*. Ypsilanti, MI: High/Scope Educational Research Foundation.

Smith, M. W., Dickinson, D. K., Sangeorge, A., & Anastasopoulos, L. (2002). *User's guide to the Early Language & Literacy Classroom Observation toolkit: Research edition*. Baltimore, MD: Brookes.

Tout, K., Zaslow, M., Halle, T., Forry, N., & Child, T. (2009). *Issues for the next decade of quality rating and improvement systems*. Retrieved from http://www.childtrends.org/Files/Child_Trends-2009_5_19_RB_QualityRating.pdf

Weiland, C., & Yoshikawa, H. Impacts of a prekindergarten program on children's.mathematics, language, literacy, executive function, and emotional skills. *Child Development*, in press.

Welsh, J. A., Nix, R. L., Blair, C., Bierman, K. L., & Nelson, K. E. (2010). The development of cognitive skills and gains in academic school readiness for children from low-income families. *Journal of Educational Psychology*, *102*(1), 43–53.

Wong, V. C., Cook, T. D., Barnett, W. S., & Jung, K. (2008). An effectiveness-based evaluation of five state prekindergarten programs. *Journal of Policy Analysis and Management*, *27*(1), 122–154.

Zakin, A. (2007). Metacognition and the use of inner speech in children's thinking: A tool teachers can use. *Journal of Education and Human Development*, *1*, 1–14.

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/programs/opre/cc/q_dot/quality_review.pdf

Zaslow, M., Martinez-Beck, I., Tout, K., & Halle, T. (Eds.). (2011). *Quality measurement in early childhood settings*. Baltimore, MD: Brookes.

Zelazo, P. D., & Müller, U. (2002). Executive function in typical and atypical development. In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development* (pp. 445–469). Oxford, UK: Blackwell.