Early Childhood Research Quarterly 60 (2022) 96–109

Contents lists available at ScienceDirect

Early Childhood Research Quarterly

journal homepage: www.elsevier.com/locate/ecresq

Content-rich instruction and cognitive demand in prek: using systematic observations to predict child gains

Michelle F. Maier[a,*], Meghan P. McCormick[a], Samantha Xia[a], JoAnn Hsueh[a], Christina Weiland[a], Abby Morales[b], Marina Boni[b], Melissa Tonachel[b], Jason Sachs[b], Catherine Snow[c]

[a] MDRC
[b] Boston Public Schools
[c] Harvard Graduate School of Education

ARTICLE INFO

Article history:
Received 12 March 2021
Revised 25 August 2021
Accepted 20 December 2021

Keywords:
Prek
Quality instruction
Classroom observations
Mathematics skills
Language skills

ABSTRACT

Currently available measures of PreK classroom quality inconsistently predict gains in children's outcomes. Extant measures may not capture the full range of instructional practices—including the degree to which children are exposed to rich content and cognitively demanding instruction—that are important for supporting the development of early language and mathematics skills. The current study leverages data from systematic observations of classrooms ($N = 51$) in public schools and community-based organizations implementing the Boston Public Schools prekindergarten program to create reliable measures of content-rich instruction and cognitive demand and explore associations between these constructs and gains in children's ($N = 378$) language and mathematics skills during PreK. Findings from descriptive analyses revealed that classrooms used content-rich and cognitively demanding practices at moderate levels, and classrooms with higher levels of intervention fidelity generally used higher-quality practices. Classrooms with higher percentages of Black students scored lower on observed cognitive demand. Results from multi-level models revealed that content-rich instruction consistently predicted gains in mathematics skills. There were stronger, positive associations between both content-rich instruction and cognitive demand and gains in mathematics skills for children who started the year with stronger mathematics skills. Neither of the constructs predicted gains in language skills. There was no consistent evidence that associations between either of these constructs and gains in mathematics and language skills varied by children's race/ethnicity, socioeconomic status, or home language. Findings demonstrate that further work to measure and provide supports for exposure to rich instructional content in PreK is warranted.

High-quality PreK during the 4-year-old year is one promising pathway for promoting more equitable child outcomes at kindergarten entry, with particularly large benefits for children from low-income, non-White, and non-native English-speaking families (Yoshikawa et al., 2013). Yet, there is considerable variation in the quality of children's PreK experiences (Pianta, Downer & Hamre, 2016), creating challenges for supporting learning and development at scale (e.g., Auger, Farkas, Burchinal, Duncan & Vandell, 2014). Complicating matters, there is no conclusive evidence about the aspects of children's learning environments that matter most for reliably generating sustained benefits for children in PreK. Current conceptualizations and measurement of PreK classroom quality may not capture the full range of practices critical for promoting children's early academic skills (Burchinal, 2018).

An emerging body of literature highlights the potential of some less studied aspects of observed classroom instructional quality—like exposure to rich content (i.e., background and world knowledge) and cognitively demanding instruction (i.e., instruction focused on promoting inferential thinking)—to predict gains in children's learning outcomes, particularly among those who begin early childhood education with lower levels of mathematics and language skills (Neuman, 2006). Relatedly, recent work has argued that PreK curricula that explicitly support a core learning domain like mathematics, language/literacy, or social-emotional skills are the field's "best bet" for boosting classroom instructional quality and gains in targeted child outcomes (Weiland, McCormick, Mattera, Maier & Morris, 2018; Yoshikawa et al., 2013). Large early educational systems like New York City have begun to implement

* Corresponding author.
E-mail address: michelle.maier@mdrc.org (M.F. Maier).

https://doi.org/10.1016/j.ecresq.2021.12.010
0885-2006/

these models at-scale (Rojas, Morris & Balaraman, 2020). Accordingly, it is important to examine how content-rich and cognitively demanding instruction are supported by—and may vary by—fidelity to those types of curricula in early learning settings. Yet, there are no known, reliable, and valid measures of these constructs in PreK that consistently predict gains in children's academic and cognitive skills. The current study aims to address this need by accomplishing 3 research objectives: 1) examining the extent to which PreK classrooms use content-rich and cognitively demanding practices and exploring whether and how these practices vary by intervention fidelity (i.e., the extent to which the curricula are implemented as intended) and by the composition of children in the classroom; 2) estimating associations between both content-rich instruction and cognitive demand and gains in children's language and mathematics skills across the PreK year; and 3) considering whether and how these associations vary for children depending on their academic skills at PreK entry, socioeconomic status (SES), race/ethnicity, and home language.

In doing so, we add to the literature in several ways. First, we explore whether we are able to use a short set of global observational items to collect reliable measures of content-rich instruction and cognitive demand. Second, we aim to establish the added value of content-rich instruction and cognitive demand for predicting children's gains in language and mathematics skills during the PreK year, over and above another widely-used indicator of classroom process quality—the Classroom Assessment Scoring System (CLASS; Pianta, La Paro & Hamre, 2008)—which does not predict child gains in the public school sample in our context (Guerrero-Rosada et al., 2021). Third, we are able to describe the curriculum used in study classrooms in detail and are further able to control for key features of intervention fidelity, including dosage, adherence, and quality (Hulleman & Cordray, 2009), which may be alternative explanations driving gains in children's outcomes. Finally, we explore the extent to which a construct like content-rich instruction—which has typically been explored in the language/literacy literature—may also be related to gains in children's mathematics outcomes. Findings stand to build evidence on features of instruction that can be intervened on to reliably support children's learning and development.

Dimensions of classroom quality and links to children's outcomes

A number of studies have identified links between the quality of children's PreK learning experiences and gains in key outcomes of interest, such as language and mathematics skills (e.g., Auger et al., 2014; (Early et al., 2007). Yet, there is no consistent definition across studies of what "high quality" entails, and little consensus on the specific aspects of quality that are consistently related to gains in children's skills. Current conceptualizations distinguish 3 basic dimensions of classroom quality that are thought to influence child outcomes (Burchinal, 2018; (Maier, Hsueh, & McCormick, 2020): 1) *structural quality* is focused on the physical features of classrooms and how they are designed and configured (e.g., teacher-student ratios; class size; classroom materials); 2) *process quality* is focused on the nature of teacher-child interactions; and 3) *instructional quality* is focused on what and how teachers teach. We hypothesize that this last dimension includes the content that children are exposed to and the pedagogy or practices teachers use to stimulate children's cognitive development and inferential thinking.

Structural quality is thought to set the stage for process and instructional quality, but it alone is not enough to promote learning and development (Yoshikawa et al., 2013). Aspects of process quality are hypothesized to be more closely linked with children's gains than structural quality due to their focus on interactions, which are more proximal to students' experiences (Tseng & Seidman, 2007). While empirical evidence supports this theory to some degree, most studies have found small and inconsistent associations between measures of PreK process quality focused on the nature of teacher-child interactions and children's outcomes (Burchinal, 2018). Measures of instructional quality focused explicitly on *what teachers teach* and *how they teach* it (e.g., Classroom Observation of Mathematics – Environment and Teaching (COEMET), Sarama, Clements, Starkey, Klein & Wakeley, 2008; Early Language & Literacy Classroom Observation Toolkit (ELLCO), Wayne, DiCarlo, Burts & Benedict, 2007) have shown larger links with child outcomes (Burchinal, 2018; Howes et al., 2008). But these associations are modest in size (Burchinal, 2018). This work highlights a need for a more nuanced understanding of classroom quality and better measures that consistently link with children's outcomes.

Diving deeper into aspects of *instructional* quality: cognitive demand

The premise behind instructional quality as a support for children's development is guided by socioecological and sociocultural learning theories that point to the importance of bidirectional interactions between children, their caregivers, and their learning environments as mechanisms that shape learning (Bodrova & Leong, 2006; Bronfenbrenner, Morris, Damon & Lerner, 2006). Repeated opportunities for engagement in guided activities and rich conversations support meaningful learning (Beck & McKeown, 2007). Consequently, we view instructional quality as encompassing both *what* is taught and *how* teachers provide these learning opportunities.

Existing measures of instructional quality tend to place more of an emphasis on *how* teachers teach, assessing teachers' use of instructional practices or activities that promote inferential thinking, such as asking children to explain, analyze, and think more deeply about ideas and engage in rich back-and-forth conversations (e.g., Bilbrey, Vorhaus, Farran & Shufelt, 2010). These practices, which challenge children to think—that is are cognitively demanding—are thought to be particularly important because they push young children to draw on their analysis, reasoning, and inferential thinking skills (Collins, 2016).

There are several instructional practices that can be viewed as cognitively demanding, such as challenging questioning, differentiated instruction, back-and-forth conversations (particularly around cognitively challenging topics), and use of sophisticated vocabulary. Activities like asking children to explain their thinking, summarize events, or provide opinions; verbally scaffolding children to help them solve a problem on their own; and engaging children in classroom discourse have demonstrated small to moderate associations with gains in children's skills, including language and literacy skills (Hamre, Hatfield, Pianta & Jamil, 2014), reading comprehension, decoding, and mathematics skills (e.g., Collins, 2016; Cook, Roggman & Boyce, 2011; Dieterich, Assel, Swank, Smith & Landry, 2006), and cognitive self-regulation (Fuhs, Farran & Nesbitt, 2013). Differentiated instruction that aims to meet children where they are developmentally is linked with preschoolers' language, literacy and mathematics skills (DeBaryshe, Gorecki & Mishima-Young, 2009). Teacher-child conversations and book-reading activities that provide opportunities to use new or sophisticated vocabulary have shown moderate associations with better language (Dickinson & Porche, 2011; Whorrall & Cabell, 2016) and mathematics skills when mathematical language is used (e.g., Purpura, Napoli & King, 2019).

Yet, associations between these kinds of cognitively demanding practices and children's skills have not been fully consistent, with variation across learning domains and observational foci. As-

sociations appear larger in studies focused on parental stimulation (d = 0.34, Dieterich et al., 2006; d = 0.19 – 38, Cook et al., 2011) compared to studies of cognitive stimulation in *classrooms* (e.g., d = 0.06 - 0.11, Fuhs et al., 2013). Further, research suggests that teachers may miss out on opportunities to provide these kinds of rich conversations and learning (Whorrall & Cabell, 2016). Thus, teachers that intentionally try to infuse cognitively demanding practices throughout the school day—such as by turning routine transitions into learning time or capitalizing on children's natural bids for attention to have an in-depth conversation about children's interests, home, or culture—may further support children's learning. Existing work is also constrained by measurement limitations, focused on the use of inferential questioning or reinforcement of vocabulary in a shared book-reading context (e.g., Wasik & Bond, 2001), whereas extant measures of classroom quality tend to capture these practices via one-off items. This limits the field's understanding of the unique role that cognitively demanding practices—as used across a range of learning activities and not just book-reading—may play in promoting children's skill development in multiple domains.

At the same time, such cognitively demanding learning opportunities may vary depending on children's skills at PreK entry and also potentially by demographic characteristics like SES, race/ethnicity, and home language. In work examining the Head Start program, researchers have found larger impacts for children who begin the program with lower scores on measures of cognitive skills, as well as Dual Language Learner (DLL) and Hispanic children (e.g., Bloom & Weiland, 2015). This research, however, also argues that the content of Head Start may be somewhat remedial and particularly focused on supporting English proficiency for children who are DLLs and were initially tested in English. Even so, it may be that Head Start programs are able to offer these children the appropriate supports, or "zone of proximal development" (Vygotsky, 1978), to learn just beyond what they already know. It is possible that supports for higher-skilled children in programs like Head Start are not as robust, or that prior work has not been able to measure the broad and comprehensive nature of cognitively demanding opportunities that support learning for these children.

Indeed, another body of work argues that when instruction is high in cognitive demand, the learning environment may be *more likely* to support the outcomes of children who have higher initial skill levels at the start of the PreK year. For example, children with stronger skills at PreK entry, who are disproportionately likely to be from higher SES families, may benefit more from talk that is inferential in nature (e.g., Reese & Cox, 1999). In line with Vygotskian theory introduced above, children who start PreK with higher levels of skills may have a stronger knowledge base that allows them to take advantage of and learn from cognitively demanding activities because their skill level is more closely matched to the demands of their learning environment. For a child with lower initial skill levels or who is learning English during PreK, some cognitively demanding activities may be beyond their abilities, even with help from a teacher. Thus, adding nuance to the range of cognitively demanding activities and opportunities that are scaffolded and differentiated in line with children's abilities may be an important aspect of assessing cognitive demand.

Diving deeper into aspects of *instructional* quality: content-rich instruction

There are a different set of extant measures of instructional quality, such as the Emerging Academic Snaphot (SNAP; Ritchie, Howes, Kraft-Sayre & Weiser, 2001) and the Narrative Record (Farran & Bilbrey, 2004), that have been collected at-scale and capture *what teachers teach* by measuring time spent in different skill domains (e.g., mathematics, literacy, science). However, *what teachers teach* goes beyond capturing indicators of time in different subject areas. It refers not only to the skill domains that are taught but also the richness of the content underlying learning activities. Neuman, Kaefer and Pinkham (2014) describe the latter construct as content-rich instruction, or the delivery of background and world knowledge as the medium through which teachers support the development of children's skills. For example, a PreK teacher might engage in a lesson to teach letter sounds by having each letter on a flashcard with a picture of a word that starts with that letter, but with no clear connection between the individual words and letter sounds, or across the various words introduced. In contrast, in a content-rich PreK classroom, content is infused into instruction by classroom activities being connected to a topical theme or focal question. The teacher could do a similar activity but have each of the letter sounds correspond to a picture or word related to a theme, such as oceans as a habitat. The letter sounds are then discussed as they relate to new vocabulary at the same time that background information about the ocean as a habitat is provided. However, even if a teacher is integrating background knowledge into instruction, classrooms are likely to vary by how rich the content is—from more superficial (e.g., "autumn") to more inferential (e.g., "living things")—by the extent to which learning activities within a day and across time are tied to a theme, by the extent to which explicit connections to the theme are made, and by the extent to which theme-specific vocabulary are used.

Background knowledge is thought to be critical for supporting student outcomes because when engaging in an activity like reading a text or completing a complex mathematics problem, the more an individual knows about the content that the text or problem is rooted in, the better and more efficient they are at comprehending and completing it; the foundational knowledge they already have helps them make connections and detract meaning (Hiebert, Goodwin & Cervetti, 2018; Recht & Leslie, 1988). Descriptive and experimental research provide evidence for this hypothesis, demonstrating that content-rich instruction in early childhood supports background knowledge in core subject areas, in turn serving as a foundation for future academic success (Neuman et al., 2014). For example, work on language development reinforces the importance of engaging children in "knowledge-building" experiences that develop vocabulary in core subject areas (e.g., mathematics, science, social studies). These results have been further supported by experimental work (Neuman et al., 2016), including studies showing that curricula using content-rich practices stand to improve children's learning outcomes, relative to typical PreK practice (Nguyen, Jenkins and Auger Whitaker, 2018). Although there is limited information about content-rich instruction in mathematics in early learning settings, research from older grades has shown that content can be a critical vehicle for supporting understanding and perceived relevance of mathematics (Zwiep and Benken, 2012). Moreover, learning does not occur in a vacuum and is mutually supported by knowledge and competencies across learning domains (Institute of Medicine & National Research Council, 2015), which suggests that mathematics instruction grounded in content may support competencies in this domain.

Content-rich instruction and cognitive demand in the Boston Public Schools prekindergarten program

Yoshikawa and colleagues (2013) have argued that the evidence for quality improvement at-scale points to implementation of domain-specific curricula that "aim to provide intensive exposure to a given content area based on the assumption that skills can be better fostered with a more focused scope" (p. 7), supported by teacher training and coaching and monitoring of children's progress. With a few notable exceptions (e.g., Boston, NYC),

few programs nationally have adopted this approach, with most choosing to implement whole-child or global curricula that aim to address all domains of child development but do not have a specified scope and sequence and do not allow for much depth of focus on any one domain(Jenkins et al., 2018). Comparison of domain-specific vs whole-child curricula suggests that the former would be better set up to support content-rich instruction and cognitively demanding practices (Weiland et al., 2018). Yet, because limited work has assessed these constructs in settings implementing these curricula, there is little evidence for this theory.

The Boston Public Schools (BPS) Prekindergarten program, however, is an example of a model working since 2007 to implement 2 evidence-based, domain-specific curricula, supporting by initial and on-going training and coaching of teachers (Weiland et al., 2018). District prekindergarten programs, as well as ten partner community-based organizations, all implement both Opening the World of Learning (Schickedanz, Dickinson & Schools, 2005), a language and literacy curriculum that includes a social-emotional skills component in each unit, and Building Blocks (Clements & Sarama, 2007), an early mathematics curriculum that also promotes language development by requiring children to explain their mathematical reasoning verbally. The PreK curriculum is culturally responsive to the diverse students that BPS serves in 3 distinct ways: 1) giving all children access to cognitively demanding tasks 2) prominently representing children's and families' diverse backgrounds, and 3) collaboratively engaging children as active agents of their learning. Teachers in both public school and community-based programs are well-compensated relative to other scaled PreK programs (Kabay, Weiland, & Yoshikawa, 2020) and the large majority have master's degree or are en route to earning one (Weiland et al., 2018). Classrooms are generally small with no more than 20 students per teacher and the school day is 6 hour and 30 minutes long. The curriculum is structured so that a large amount of time is spent in small, structured, free-choice centers and there are opportunities for small group instruction daily (McCormick et al., 2020). Prior work examining model implementation has shown moderate to high levels of fidelity (McCormick et al., 2020). Given this fine-grained information about fidelity to the model, questions remain about the extent to which the approach does achieve its goal of improving content-rich and cognitively-demanding practice and how those practices translate into gains for students.

The current study

The current study aims to build on the existing PreK classroom quality literature by leveraging data from the BPS prekindergarten program implemented in public schools and community-based organizations. The study will answer 3 research questions:

1 To what extent do classrooms delivering the BPS prekindergarten model use content-rich and cognitively demanding practices and how do these constructs vary by intervention fidelity and by the composition of children in the classroom?
2 Does exposure to content-rich and cognitively demanding practices predict gains in children's language and mathematics skills during the PreK year?
3 Do these associations vary for children who start the academic year with weaker vs stronger language and mathematics skills and by children's SES, race/ethnicity, and DLL status?

Taken together, findings will shed light on observable characteristics that may not be captured in existing measures of classroom quality and determine whether they are potential targets for future measurement and intervention work in order to support gains for children entering early childhood settings with varying levels of skills and demographic characteristics.

Method

Participants and setting

The sample for the current study consists of 378 children enrolled in prekindergarten during the 2016 - 2017 academic year, recruited from 51 classrooms within 20 public elementary schools and 10 community-based programs (referred to as CBOs hereafter) all implementing the BPS prekindergarten model during the academic year. Students in the current study were enrolled in classrooms with observational data and participated in assessments of their language and mathematics skills in the fall and spring of the PreK year.

In public school settings, the BPS prekindergarten program is free, runs full-day, and is open to any age-eligible child for the academic year (though there is more demand than supply). All BPS prekindergarten teachers meet the same requirements and receive the same compensation as K-12 teachers and are required to have an early childhood (preschool to grade 2) license from the Massachusetts Department of Elementary and Secondary Education and have or be working towards a master's degree in education. The district recently partnered with a subset of CBOs offering center-based PreK to implement this model as well. At least one teacher in every CBO classroom implementing the BPS prekindergarten model had a minimum of a BA in early childhood education or a related field. Teachers in partnering CBOs received a pay boost to ensure parity with the entry-level salary of teachers in public school settings. All classrooms included in the current study implemented the BPS *Focus on K1* curriculum, with implementation supported through district-provided training and coaching.

The demographic characteristics of the child and teacher samples are presented in Table 1. As illustrated in the top panel of the table, the majority of the child sample was eligible for free- or reduced-price lunch, and the children were diverse with respect to racial/ethnic background and parental education, among other characteristics. We coded all students at the CBO programs as eligible for free- or reduced-price lunch, due to income eligibility for those slots coupled with reports from the CBO providers. Teachers also come from diverse racial/ethnic backgrounds (13% Hispanic, 43% White, 26% Black, 9% Asian, and 9% other race or biracial) with an average of 14.83 (SD = 8.86) years of teaching experience. Compared to the sample, students in the general population of BPS public prekindergarten were more likely to be Hispanic (28% of the sample compared to 38% of the broader population) and less likely to be Asian (14% of the sample, compared to 9% of the broader population).

Procedures

School and classroom recruitment

In the summer of 2016, we randomly selected 25 public schools to participate in the study from the 76 district schools offering the public prekindergarten program. 21 agreed. We used 1 school as a pilot school for measure development and the remaining 20 schools made up the public school sample. We then selected 10 of the 11 CBOs implementing the BPS prekindergarten model to participate in the study and all agreed. As mentioned, the CBOs were connected to the BPS Department of Early Childhood (DEC) and were receiving training and coaching to implement the BPS prekindergarten model during the 2016 – 2017 year. We asked all teachers assigned to general education or inclusion classrooms in each of the 20 public schools to participate in the study in the fall of 2016. We randomly selected 1 classroom serving 4-year-old children within each CBO to participate. 96% ($N = 51$) of teachers agreed, and there was no attrition across the year.

Table 1
Descriptive characteristics of child, teacher, and classroom sample.

Characteristic	Mean or%	SD	Percent missing
Child characteristics Race/ethnicitv (%)			
Hispanic	28.38	–	26.26
White	22.02	–	26.26
Black	31.03	–	26.26
Asian	14.06	–	26.26
Other race	4.51	–	26.26
Female (%)	50.26	–	0.00
Child age on September 1st, 2016 (fall of PreK)	4.50	0.30	0.00
Eligible for free or reduced price lunch (%)	66.93	–	0.00
Dual Language Learner (%)	48.68	–	0.00
Days between Fall and Spring assessments (in years)	0.52	0.07	0.00
PPVT raw score in Fall of PreK	72.20	27.38	2.91
PPVT raw score in Spring of PreK	85.57	26.50	5.56
WJAP raw score in Fall of PreK	12.04	5.14	3.44
WJAP raw score in Spring of PreK	14.98	4.96	5.82
REMA t score in Spring of PreK	36.37	6.61	5.56
Teacher/classroom characteristics Teacher race/ethnicity (%)			
Hispanic	13.04	–	9.80
White	43.48	–	9.80
Black	26.09	–	9.80
Asian	8.70	–	9.80
Other race	8.70	–	9.80
Years of teaching experience	14.83	8.86	7.84

Note: $N = 378$ PreK students; $N = 51$ PreK classrooms.

Recruitment of children

Eighty-one percent of all children in participating classrooms had parent consent to participate in the study. The team randomly selected 50% (~6 – 10 per classroom) of consented children to participate in data collection. The children in this group were representative of the broader population of children with written consent to participate.

Direct assessments

The field-based research team assessed children in prekindergarten during fall 2016 (October 1 through December 12) and spring 2017 (April 5 through June 16). All child assessors were trained to reliability prior to collecting data. Before beginning the study battery, assessors used the Pre-language Assessment Scale (preLAS; (Duncan & DeAvila, 1998) to determine the administration language for a subset of assessments (Barrueco, López, Ong & Lozano, 2012). Of the 378 children in the sample, 43 (11%) completed a subset of assessments in Spanish in fall 2016, and 15 (4%) completed assessments in Spanish in spring 2017. There were $N = 363$ children who completed the assessments in fall, $N = 356$ who completed assessments in spring, and $N = 341$ who completed assessments at both time points.

Classroom observations

We provide a description of procedures for live and videotaped classroom observations in the main text and further details in Appendix B.

Live classroom observations

In winter 2017, trained instructional coaches from the BPS district observed each participating classroom for two 2-hour blocks of academic instruction (*Mean* = 106 minutes, *SD* = 29). Coaches only completed fidelity observations in classrooms they did not normally coach. Observations focused on the full classroom including the lead teacher, the children, and any assistant teachers and other adults.

Fifteen BPS instructional coaches trained as observers participated in a 3-day training in January 2017 to learn how to rate classrooms and teachers on different indicators of intervention fidelity for each component of the *Focus on K1* curriculum. Coaches were majority female (80%) and diverse with respect to race/ethnicity (33% White, 33% Black, 20% Hispanic, 13% Asian). All were former early childhood or elementary school teachers. As part of the training, observers also learned how to rate classrooms on a new measure created to capture global indicators of instructional quality, regardless of the curricular components observed (see shortened set of items in Table 2 and full measure with anchors and descriptions in Appendix C). During the data collection period (February – May 2017), the team double-coded 20% of observations to assess interrater reliability, showing high levels of agreement, with 89% of fidelity double codes being reliable "within 1″ (aligned with reliability standards on observational measures like the CLASS).

Videotaped classroom observations

On different days than the live observations, we also collected 2 videotaped observations of instruction (*Mean* = 3.16 hour of total time across the observations, *SD* = 0.83, min = 2.21 hour, max = 4.62 hour) during winter 2017. We coded observations using the CLASS (Pianta et al., 2008) and an adapted version of the Individualizing Student Instruction (ISI; Connor et al., 2009) tools. We used CLASS domains as covariates in our main analyses and robustness checks (the 3 domains together in our main analysis and each domain separately in robustness checks) and data from the ISI for robustness checks only.

All coders participated in a 2-day CLASS training led by a certified trainer and then established reliability on a set of master codes created by the developers. Coding of each videotape started once the instructional time began. As recommended by the measure's protocol (Pianta et al., 2008), coders used cycles of 20 minutes for observing and 10 minutes for scoring, which they repeated up to 4 times for each videotape. We averaged scores across the 4 segments and then across the 2 observations to generate overall scores for each classroom. We double-coded 20% of the observations to assess interrater reliability. The final ICCs representing interrater reliability for the 3 domains were 96% for Emotional Support, 94% for Classroom Organization, and 88% for Instructional Support. We also conducted drift checks wherein coders had to code a master tape every 3 weeks to ensure they were still reliable.

Table 2
Results of exploratory factor analysis with two-factor solution.

Item	Factor loadings: varimax rotation	
	Construct 1 cognitive demand	Construct 2 content-rich instruction
Learning opportunities in this classroom cognitively demanding	0.79	0.42
Classroom capitalizes on learning opportunities for children	0.78	0.49
Teacher talks to children in ways that encourage them to expand on or think more deeply about ideas	0.75	0.44
Relevant vocabulary and rich academic language used and clearly defined throughout the observation	0.71	0.48
Teacher uses differentiated learning strategies to make the curriculum accessible to a range of children	0.51	0.26
Classroom culture to discuss and explicitly demonstrate diversity	0.50	0.10
Evidence of the theme/focal question in this classroom instructional time	0.12	0.93
Rich content delivered on the theme/focal question	0.43	0.77
Teacher made connections between activities to deepen children's understanding of the theme/focal question	0.48	0.70
Teacher defined theme-specific vocabulary words	0.29	0.63
Evidence of the theme in classroom materials, including materials within centers and students' work on walls	0.26	0.63
Teacher connects or links activities to the curriculum unit or book in explicit and intentional ways	0.28	0.60
Abstract content delivered on the theme/focal question	−0.21	−0.04
Construct alpha	0.90	0.90

Note: $N = 51$ classrooms. Items were rated on a scale from 1 to 5, with a higher score indicating the teacher performed better on that item. Factor loadings considered in measurement creation are shaded.

Administrative data

We accessed administrative records from the BPS district on children's demographic characteristics (race, ethnicity, birthdate, eligibility for free/reduced price lunch, home language, sex), history of enrollment in the BPS prekindergarten program, and current classroom and school membership. We lacked these data on students from CBOs and thus created these indicators using a combination of parent reports (race/ethnicity, birthdate, home language) and broader information on income eligibility for CBO slots.

Measures

Intervention fidelity to the BPS Focus on K1 *model: dosage, adherence, and quality*

The research team worked closely with the BPS DEC to create an observational tool to assess fidelity to the integrated *Focus on K1* model—described above— in prekindergarten classrooms. We published information describing the creation of this tool and our assessment of its reliability and validity (McCormick et al., 2020). In line with recommendations from Hulleman and Cordray (2009), we used this tool to measure 3 dimensions of intervention fidelity – dosage, adherence, and quality. We used these measures of intervention fidelity to conduct robustness checks. See Appendix D for an overview of the *Focus on K1* curricular components.

Implementation dosage. Dosage captured the number of curricular components observed and the amount of time spent doing them.

Adherence to the curriculum. Adherence items captured whether particular aspects of each curricular component were implemented as intended. We coded adherence to the curriculum as the proportion of adherence items that the teacher was observed to implement within each component. We then averaged across all of the observed components to calculate a total adherence score (which was a percentage ranging from 0 to 100).

Quality of curricular implementation. Quality items captured the manner by which the curricular components were delivered and whether particular instructional practices were used. These items included a detailed set of anchors and descriptors using a 5-point Likert scale where 1 = low quality and 5 = high quality. We calculated the average quality score for each teacher across all quality items in curricular components that were observed.

Scoring. We averaged across both observations to create classroom-level fidelity scores. We report on descriptive statistics for the dimensions of adherence and quality and use those variables in our robustness checks described in Appendix A. Finally, we calculated the number of days between September 1 and the fidelity observations, averaged across the 2 observations, and included that as a covariate in the predictive models.

Content-rich instruction and cognitive demand

Recognizing that it would be impossible to observe the same curricular components in every observation or classroom, we also developed a set of *global items* aimed at capturing content-rich instruction and cognitively demanding practices. We used these items to rate every classroom on a consistent set of items using a 5-point Likert scale (1 = low, 5 = high), regardless of the curricular components that were observed. Observers rated the extent to which instruction was both rich in content and vocabulary and cognitively demanding (see Table 2 for a list of items; full measure included in Appendix C). Although these global items were inspired by extant observational tools (Pianta et al., 2008) and prior fidelity measures used in BPS (Yudron, Weiland & Sachs, 2016), the content was driven primarily by the overarching goals of the BPS approach, including the promotion of rich vocabulary, conceptual knowledge, and an understanding of abstract, complex ideas. After conducting analyses to assess the items' psychometric properties, we averaged the items within domains to create 2 separate con-

structs representing cognitive demand ($N = 6$ items) and content-rich instruction ($N = 6$ items).

Classroom process quality

We measured global classroom process quality using the Classroom Assessment Scoring System (CLASS) PreK (Pianta et al., 2008). CLASS measures 3 domains of teacher-child interactions: Emotional Support, Classroom Organization, and Instructional Support. All dimensions are directly scored on a 7-point scale, where a score of 7 represents high quality except for negative climate which is reverse-coded. The CLASS and these 3 constructs show good psychometric reliability and validity in the literature, and prior studies examining associations between quality and children's outcomes have used this same 3-factor structure (Burchinal et al., 2014). We included all 3 domains as one block of covariates. We also examined the sensitivity of these results to models fit separately for each CLASS domain (discussed in robustness check section and Appendix A).

Language skills

We used the *Peabody Picture Vocabulary Test IV* (PPVT IV; Dunn & Dunn, 2007) to assess children's receptive language skills in the fall and spring. The PPVT IV is a nationally normed measure used widely in diverse samples of young children. The test has excellent split-half and test–retest reliability estimates, and strong qualitative and quantitative validity properties (Dunn & Dunn, 2007). We assessed all children on the PPVT (regardless of whether they passed the PreLAS language screener) in order to describe an equivalent measure of receptive language skills in English across the sample.

Mathematics skills

We assessed children's mathematics skills using both the *Woodcock Johnson Applied Problems* (WJAP; Woodcock, Mather, McGrew & Wendling, 2001) and the *Research-based Early Mathematics Assessment* (REMA; Clements, Sarama & Liu, 2008). Children who did not pass the PreLAS screener ($N = 43$ in fall and $N = 15$ in spring) were assessed using the Spanish versions (i.e., the Batería III Woodcock Muñoz and the Spanish translation of the REMA). We combined scores from the English and Spanish assessments together for the sample. Full details on the reliability and validity of these measures are included in Appendix E.

The *WJ/WM Applied Problems* assessment is a numeracy and early mathematics measure that requires children to perform calculations to analyze and solve arithmetic problems. It has demonstrated good evidence of reliability and validity in prior work. We present results using the raw score of the measure. The WJ Applied Problems subtest has been criticized by some mathematics experts because it is not particularly sensitive in the early childhood years, skips quickly to difficult items, and does not include geometry (Weiland et al., 2012). Accordingly, in spring of prekindergarten, we also used the REMA to assess children's early mathematics skills. The REMA is a hands-on, one-on-one assessment that measures core mathematical abilities of children ages 3 – 8 and has demonstrated good psychometrics (Clements et al., 2008). The REMA was only collected in the spring because the updated version of the assessment was not yet available. We used the WJAP as the fall baseline level of mathematics in all predictive models.

Child characteristics from administrative and assessment data

Using administrative data, we created a series of indicators to describe children's race/ethnicity (Black, Hispanic, Asian, or Other Race/Ethnicity (including mixed-race children)) with the reference group as White. We used similar indicators to describe eligibility for free or reduced-price lunch and sex (1 = female; 0 = not female). We set a dummy variable for DLL equal to 1 if the parent reported that there was a language other than English ever spoken at home and 0 otherwise. We used the child's birthdate to calculate child age on September 1, 2016. We also included the number of days between the fall and spring assessments as a covariate.

Analytic approach

Missing data

Overall, there was a relatively low amount of missing data. All students had complete data on child-level information provided by the school district. Missingness on variables used in analyses ranged from 0% to 6%. There was limited evidence that data were systematically missing. Accordingly, we chose to use listwise deletion to fit our models when answering the key research questions of interest. However, as a robustness check we did use multiple imputation similar to how we have done on a number of other studies with this sample (and different outcome measures; see (McCormick et al., 2020)(McCormick et al., 2021) . More information on treatment of missing data is included in Appendix A.

Factor analyses

To examine the construct validity of our measures of cognitive demand and content-rich instruction, we conducted an exploratory factor analysis (EFA) with varimax rotation using the full sample of 51 prekindergarten classrooms (both public school and CBO) that participated in live observations. We fit this model using the classroom-level scores averaged across the 2 observation periods. Results from both the scree plot and the rotated factor solution suggested that a 2-factor solution was the best fit to the data. 12 of 13 items loaded onto one of the factors at 0.5 or above. There was 1 item (denoted in Table 2) that did not load onto either factor and was excluded from the creation of these constructs. In addition, there were some instances where an item loaded onto 1 factor at 0.7 or above and then loaded onto the other factor at 0.42 to 0.49. However, because it was clear that these items had a high loading on 1 factor and a substantially lower loading on the other factor, we were not concerned about issues with double-loading items. Table 2 lists the loadings for the rotated 2-factor solution. The 2-factor solution fit the data better than a 3- or 4-factor solution.

After reviewing the items within each factor, we labeled the first factor cognitive demand ($\alpha = 0.90$) and the second factor content-rich instruction ($\alpha = 0.90$). Cognitive demand included items about the extent to which the learning opportunities provided required higher-order or strategic and extended thinking (compared to opportunities that were more rote or basic in nature) or focused on different aspects of diversity (e.g., language, cultures, gender), whether the teacher appeared to capitalize on opportunities to extend children's learning in intentional ways, and the frequency with which rich vocabulary and language was used and defined. Content-rich instruction was comprised of items tapping into the extent to which there was evidence of background knowledge (e.g., families, things that grow, habitats) being presented via a unit theme or focal question throughout different learning opportunities and in the vocabulary used, the extent to which teachers made connections between different activities and referenced that background knowledge in intentional ways, and the extent to which the background knowledge provided was rich in nature—that is promoted conceptual thinking.

We created cognitive demand and content-rich instruction scores by calculating the average rating of the non-missing items within each factor for each classroom in the sample. We examined bivariate correlations with the CLASS domains, finding low correlations between cognitive demand and CLASS emotional support ($r = 0.23$) and classroom organization ($r = 0.28$), and modest correlations between content-rich instruction and CLASS emo-

tional support ($r = 0.08$) and classroom organization ($r = 0.08$). Correlations for cognitive demand ($r = 0.33$), as well as content-rich instruction ($r = 0.23$), with instructional support were slightly larger than for the other CLASS domains. Correlations between these practices and the 3 dimensions of the instructional support were low but higher for cognitive demand ($rs = 0.29–0.31$) than content-rich instruction ($rs = 0.20–0.23$). Cognitive demand and content-rich instruction constructs were moderately correlated with one another ($r = 0.71$).

Descriptive analysis

We first examined classrooms' scores on the cognitive demand and content-rich instruction measures using descriptive statistics. We then examined bivariate correlations between cognitive demand and content-rich instruction and information on intervention fidelity and explored whether these constructs varied by intervention fidelity and by the composition of children in the classroom in terms of their race/ethnicity, DLL status and eligibility for free or reduced-price lunch (our proxy for SES).

Multi-level modeling

We used multi-level modeling to answer our second and third research questions about whether exposure to content-rich and cognitively demanding practices predicted children's gains and whether these associations varied by children's language and mathematics skills at the beginning of the prekindergarten year. Because children in the sample were nested in classrooms and schools/CBOs, we fit unconditional means models to partition the variance at each relevant level (i.e., school/center, classroom, child). The 3-level model with random intercepts for classrooms and schools was the best fit to the data across outcomes (Snijders & Bosker, 2011). Eq. (1) below illustrates the model we fit to answer our second research question, examining associations between content-rich instruction and cognitive demand and gains in children's language and mathematics skills.

$$Y_{ijk} = \beta_0 + \beta_1 ContentRichInstruct_{jk} + \beta_2 CogDemand_{jk} + \gamma_{ijk} + \alpha_{jk} + \mu_{jk} + \zeta_k + \varepsilon_{ijk}, \quad (1)$$

As summarized here, we regressed each outcome measured in the spring of prekindergarten for student i, in classroom j and school k (Y_{ijk}) on both content-rich instruction and cognitive demand, together as one block. In Eq. (1), γ_{ijk} is a vector of student-level covariates, α_{jk} is a vector of classroom-level covariates, μ_{jk} and ζ_k denote random intercepts for classrooms and schools, and ε_{ijk} is a residual error term. We made the decision to include content-rich instruction and cognitive demand as predictors together in the model because the correlation between the measures was moderate ($r = 0.71$) and we wanted to understand their unique association with outcomes, net of the other. We grand-mean centered all continuous variables.

Finally, we tested our third research question—considering variation in associations between indicators of quality and gains in children's outcomes by fall skill level as well as SES, race, and ethnicity—by fitting separate models for each moderator and domain of quality (moderators: children's academic skills in the fall of the academic year, SES, 5 categories of race/ethnicity used as covariates in earlier models, DLL status), adjusting for covariates, including the other theorized quality construct.

Results

Descriptive analysis

We summarize descriptive findings in Table 3. CLASS scores were lower than those reported in a previous study of classroom quality in the BPS prekindergarten program (Weiland & Yoshikawa, 2013). With respect to intervention fidelity, there were moderately high levels of adherence (63.6%) and moderate levels of quality of implementation (mean = 3.25, $SD = 0.56$) across observed curricular components. We observed about 75% of the curriculum components in each classroom (mean = 8.47, $SD = 2.58$) with variation attributed to the frequency with which some components (like Centers, Read Aloud) took place compared to others.

Research Question 1. To what extent do classrooms implementing the BPS prekindergarten model use content-rich and cognitively demanding practices and how do these constructs vary by intervention fidelity and the composition of children in the classroom?

As described in Table 3, we found that classrooms used moderate levels of content-rich (mean = 3.10, $SD = 0.79$) and cognitively demanding (mean = 2.89, $SD = 0.75$) practices. Exploratory analyses taking intervention fidelity into account revealed fairly large and statistically significant associations between both content-rich instruction and quality of curricular implementation ($r = 0.68$, $P < 0.001$) and cognitive demand and quality of curricular implementation ($r = 0.74$, $P < 0.001$). There were smaller—yet still moderately sized—statistically significant associations between content-rich instruction and adherence to the curriculum ($r = 0.44$, $P < 0.01$) and cognitive demand and adherence to the curriculum ($r = 0.39$, $P < 0.01$). We then explored whether content-rich instruction and cognitive demand were generally similar or varied depending on the composition of the children in the classroom. We found a negative correlation between the proportion of Black children and cognitive demand, which was statistically significant at $P < 0.10$ ($r = -0.25$, $P < 0.10$). There were no other correlations greater than 0.20 between content-rich instruction or cognitive demand and the other classroom demographic characteristics (i.e., racial/ethnic categories, SES, or DLL status).

Research Question 2. Does exposure to content-rich and cognitively demanding practices predict gains in children's language and mathematics skills across the prekindergarten year?

Findings across all models are summarized in Table 4. For statistically significant results we present results from full models – with covariates – in Table 5. Content-rich instruction was positively associated with gains in children's mathematics skills (measured using the WJAP) in prekindergarten ($\gamma = 1.20$, $SE = 0.34$, $P < 0.001$, std. association = 0.24) while cognitive demand was negatively associated with gains in mathematics skills during this same time ($\gamma = -0.97$, $SE = 0.34$, $P < 0.01$, std. association = −0.20). In other words, holding all other factors constant, including the level of cognitive demand, children who were exposed to 1 standard deviation more content-rich instruction demonstrated larger gains in mathematics skills—of about 0.24 SDs (or about 2.5 months of learning in mathematics; Hill, Bloom, Black & Lipsey, 2008)—than children exposed to the mean level of content-rich instruction. Holding all covariates constant, including content-rich instruction, children exposed to 1 SD more cognitive demand experienced smaller gains in mathematics—of about 0.20 SDs—than children exposed to the mean level of cognitive demand. Further, models demonstrated null associations between both of these predictors and gains in children's language skills and spring mathematics skills assessed with the REMA in spring.

Research Question 3. Do these associations vary for children who start the academic year with weaker vs stronger language and mathematics skills and by children's SES, race/ethnicity, and DLL status?

All interaction results are presented in Table 4. Results from the full models with statistically significant interactions – including covariates – are in Table 5. There was a statistically significant interaction between content-rich instruction and fall skills in the model predicting spring REMA scores ($\gamma = 0.12$, $SE = 0.06$, $P < 0.05$). Children who entered school with stronger levels of math-

Table 3
Descriptive statistics and bivariate correlations for measures of classroom quality and intervention fidelity†.

Variable of interest		Mean	SD	1	2	3	4	5	6	7
1.	Content-rich instruction	3.10	0.79	–						
2.	Cognitive demand	2.89	0.75	0.71 ***	–					
3.	CLASS emotional support	5.51	0.59	0.08	0.23	–				
4.	CLASS classroom organization	5.36	0.58	0.08	0.28 *	0.85***	–			
5.	CLASS instructional support	3.19	0.63	0.23	0.33 *	0.63 ***	0.68 ***	–		
6.	Quality of implementation	3.25	0.56	0.68 ***	0.74 ***	0.23	0.22	0.33 *	–	
7.	Adherence to curriculum	0.64	0.11	0.44**	0.39**	−0.15	−0.21	−0.04	0.57***	–
8.	#curricular components observed	8.47	2.58	0.08	0.05	−0.35 *	−0.20	−0.12	0.00	0.12

Note: $N = 51$ classrooms. Sample has limited missing data.
*** $P < 0.001$.
** $P < 0.01$.
* $P < 0.05$.
† $P < 0.10$.

Table 4
Associations between instructional practices and gains in students' language and math skills in PreK.

Fixed effects	Language skills (PPVT)			Math skills (WJAP)			Math skills (REMA)		
	y	SE	Std.association	y	SE	Std.association	y	SE	Std.association
Main effects model	0.26	2.07	0.01	−0.97 **	0.34	−0.20	−1.08	0.68	−0.16
Cognitive demand									
Content-rich instruction	−0.16	2.11	−0.01	1.20 ***	0.34	0.24	0.80	0.70	0.12
Emotional support	−1.18	3.90	−0.04	−1.63*	0.66	−0.33	−1.32	1.28	−0.20
Classroom observation	4.05	4.35	0.15	1.59*	0.70	0.32	1.98	1.43	0.30
Instructional support	−2.91	2.25	−0.11	0.11	0.38	0.02	−0.76	0.75	−0.12
Interaction model for cognitive demand	0.45	2.10	0.02	−0.97 **	0.34	−0.20	−1.03	0.68	−0.16
Cognitive demand									
Cognitive demand x baseline skill	0.08 †	0.05	0.00	0.00	0.05	0.00	0.15 *	0.07	0.02
Interaction model for content-rich instruction	−0.24	2.13	−0.01	1.21 ***	0.35	0.24	0.71	0.70	0.11
Content-rich instruction									
Content-rich instruction x baseline skill	0.04	0.04	0.00	−0.02	0.04	0.00	0.12 *	0.06	0.02

Note: $N = 335$ students and $N = 50$ classrooms. All models also adjust for Fall assessments, time between assessments, student age, free/reduced price lunch status, sex, dual-language learner status, whether student was in a CBO or not, and average number of days between September 1 and fidelity observations. Interacted models also adjust for the three domains of the CLASS (emotional support, classroom organization, instructional support).
*** $P < 0.001$.
** $P < 0.01$.
* $P < 0.05$.
† $P < 0.10$.

ematics skills and who experienced high levels of content-rich instruction demonstrated higher mathematics scores during spring of prekindergarten than children with similar fall skills who experienced lower levels of content-rich instruction (see Fig. 1A). There was also a statistically significant interaction between cognitive demand and fall skills in the model predicting the REMA ($\gamma = 0.15$, $SE = 0.07$, $P < 0.05$). As illustrated in Fig. 1B, children who began prekindergarten with lower levels of mathematics skills and experienced lower levels of cognitive demand showed stronger mathematics skills on the REMA than children who began school with similar levels of academic skills and experienced higher levels of cognitive demand. Finally, in the models predicting gains in language skills, we also found an interaction between fall skills and cognitive demand ($\gamma = 0.08$, $SE = 0.05$, $P < 0.10$). Further inclusion of an additional set of controls (discussed in Appendix A) yielded a parameter estimate of similar-sized magnitude ($\gamma = 0.10$) on this interaction that was statistically significant at $P < 0.05$ (see more below). Probing the interaction revealed that children who entered prekindergarten with higher levels of language skills and experienced higher levels of cognitive demand demonstrated larger gains on the PPVT during prekindergarten than children who entered school with similar levels of language skills and were ex-

Table 5
Full multi-level models predicting gains in prek from content-rich instruction, cognitive demand, and interactions between constructs and fall skills.

	Lang skills (PPVT)		Math skills (WJAP)		Math skills (WJAP)		Math skills (WJAP)		Math skills (REMA)		Math skills (REMA)	
Fixed effects	γ γ	SE 2.58	γ	SE 0.43	γ	SE 0.43	γ	SE 0.43	γ	SE 0.78	γ	SE 0.78
Intercept	88.86 ***		15.62 ***		15.62 ***		15.60 ***		37.78 ***		38.05 ***	
Covariates												
Fall level of the outcome (or skill)	0 69 ***	0.04	071 ***	0.04	0.71 ***	0.04	0.71 ***	0.04	0 72 ***	0.06	0 74 ***	0.05
Free/reduced price lunch eligible	−4.06	2.48	−0.01	0.47	−0.01	0.47	−0.01	0.47	−0.40	0.69	−0.47	0.69
Female	−0.53	1.62	−0.08	0.31	−0.07	0.31	−0.07	0.31	−0.17	0.45	−0.25	0.46
Dual Language Learner	−1.74	2.12	0.47	0.40	0.47	0.40	0.48	0.40	1.23 *	0.59	1.14 f	0.59
Asian	2.46	3.31	−0.54	0.60	−0.53	0.60	−0.54	0.60	−0.47	0.95	−0.55	0.95
Black	0.80	2.97	−1.10 f	0.56	−1.10 f	0.56	−1.09 f	0.56	−2.27 **	0.86	−2.28 **	0.86
Hispanic	−2.12	2.90	−1.36 *	0.56	−1.36 *	0.56	−1.34 *	0.56	−2.11 *	0.82	−2.21 **	0.83
Other race	6.00	3.96	0.35	0.76	0.35	0.76	0.37	0.76	−0.07	1.12	−0.12	1.12
Community-based PreK	1.76	4.64	0.53	0.73	0.52	0.74	0.51	0.73	−0.49	1.52	−0.56	1.51
Child age at school year start	1.33	2.82	−0.43	0.55	−0.43	0.55	−0.43	0.55	0.17	0.81	0.20	0.81
Time between assessments	12.83	14.28	2.64	2.65	2.65	2.67	2.61	2.66	6.40	4.18	5.80	4.17
Days between 9/1 and fidelity obs. CLASS domains	−0.07	0.10	−0.04 **	0.02	−0.04 **	0.02	−0.04 **	0.02	−0.02	0.03	−0.02	0.03
Emotional support	−2.32	4.00	−1.63 *	0.66	−1.62 *	0.66	−1.61 *	0.66	−1.47	1.27	−1.38	1.27
Classroom organization	5.20	4.45	1.59 *	0.70	1.59 *	0.70	1.57 *	0.70	2.19	1.43	2.12	1.43
Instructional support Global quality domains	−2.92	2.27	0.11	0.38	0.11	0.38	0.11	0.38	−0.91	0.75	−0.87	0.75
Cognitive demand	0.45	2.10	−0.97 **	0.34	−0.97 **	0.34	−0.98 **	0.34	−1.03	0.68	−1.02	0.68
Content-rich	0.04	2.14	1.20 ***	0.34	1.20 ***	0.35	1.21 ***	0.35	0.81	0.70	0.71	0.70
Interaction terms												
Cognitive demand x baseline skill	0.08 †	0.05	–	–	0.00	0.05	–	–	0.15 *	0.07	–	–
Content-rich x baseline skill	–	–	–	–	–	–	−0.02	0.04	–	–	0.12 *	0.06

Note: $N = 335$ students and $N = 50$ classrooms. All models also adjust for Fall assessments, time between assessments, student age, free/reduced price lunch status, sex, dual-language learner status, whether student was in a CBO or not, and average number of days between September 1 and fidelity observations, and the three domains of the CLASS (emotional support, classroom organization, instructional support). Given space constraints, we do not present null models. These results are available by request.

*** $P < 0.001$.
** $P < 0.01$.
* $P < 0.05$.
† $P < 0.10$.

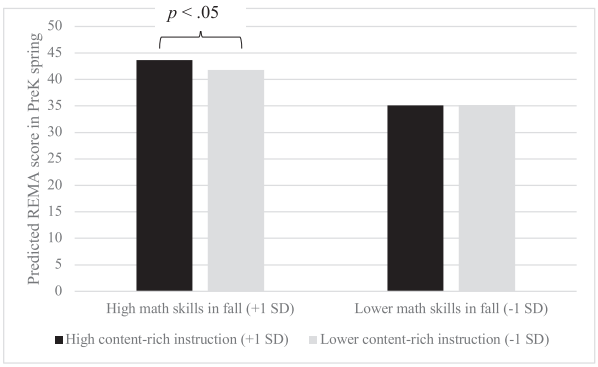
posed to lower levels of cognitive demand. We treat this finding as exploratory given the inconsistency in the p value, but it does suggest that cognitive demand may act similarly for language and mathematics skills, in predicting gains for children who begin the year with stronger skills. We did not find any consistent evidence that associations between either content-rich instruction or cognitive demand and gains in skills varied by students' eligibility for free or reduced price lunch, race/ethnicity, or DLL status.

Robustness checks

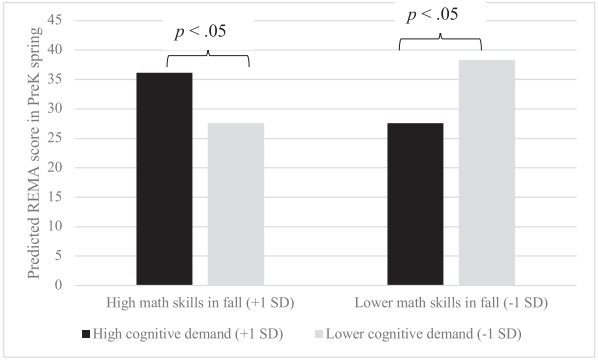
We considered a number of robustness checks (see Appendix A) including: multiple imputation to handle missing covariates; fitting separate predictive models for content-rich instruction and cognitive demand; and examining sensitivity of results to a range of alternative explanations (amount of time children spent in language and mathematics instruction, the number of curricular components classrooms were observed to implement [i.e., intervention dosage], intervention adherence, and overall quality of implementation). The association between content-rich instruction and gains in mathematics skills (assessed on the WJAP) was fully robust and varied little in magnitude across all of our checks ($N = 6$). In contrast, the negative association between cognitive demand and gains in the WJAP was no longer significant and had a decreased magnitude in 2 out of 6 checks. As such, we have more confidence in the predictive ability of content-rich instruction than cognitive demand. Although the interactions were not all robust to multiple imputation at $P < 0.05$, prior work has shown that multiple imputation can introduce more bias and error than complete case analysis when the level of missingness is low relative to sample size (Gelman and Hill, 2006). There is value in considering these interactions in an exploratory way, as they were robust to alternative models and inclusion of conceptually meaningful covariates.

Discussion

This paper aimed to build evidence on the reliability and predictive ability of 2 aspects of PreK classroom instructional quality—content-rich instruction and cognitive demand—hypothesized to be associated with gains in children's language and mathematics skills. We were able to create reliable measures of these constructs by collecting systematic observations of classrooms implementing domain-specific, evidence-based curricula designed to support these teacher practices. Descriptive findings revealed that classrooms used content-rich and cognitively demanding practices at moderate levels (on a 5-point scale) and that our measures were sensitive enough to detect variation in these constructs across study classrooms. Importantly, we found evidence that classrooms with higher proportions of Black students scored lower on our measure of cognitive demand, compared to classrooms with lower proportions of Black students. Content-rich instruction was con-



a



b

Fig. 1. (A). Interaction of fall mathematics skills and content-rich instruction predicting spring REMA scores in PreK, (B). interaction of fall mathematics skills and cognitive demand predicting spring REMA scores in PreK.

sistently associated with gains in children's mathematics skills—measured using the Woodcock Johnson Applied Problems—at a magnitude of about a quarter of a standard deviation, whereas findings related to main effects of cognitive demand and children's gains were inconsistent as were associations between content-rich instruction and gains in language skills. There was no evidence that these associations varied by students' race/ethnicity, SES, or DLL status.

Support of content-rich instruction and cognitive demand via curriculum

Specific to our descriptive analysis, our results show that stronger implementation of domain-specific curricula focused on both language/literacy and mathematics can support higher levels of content-rich and cognitively demanding instruction. Classrooms with higher levels of intervention fidelity to BPS's *Focus on K1* curriculum were observed using more content-rich and cognitively demanding practices. For example, compared to teachers who were observed adhering to about 40% of observed curricular components (i.e., 1 standard deviation below the mean), teachers who implemented curricular components with 80% adherence (1 standard deviation above the mean) scored about half of a standard deviation higher on our measure of content-rich instruction and about a third of a standard deviation higher on our measure of cognitive demand. This is a notable finding because it identifies curricula as a clear approach for programs to hone in on to enhance these features of instructional quality. In addition, given the difficulty of collecting fine-grained measures of intervention fidelity for a large number of classrooms, it may be more feasible to consider approaches like ours—using global observational measures—to assess content-rich instruction and cognitive demand at-scale.

Recent updates in standards for early childhood education highlight the importance of using curricula to support both content-rich and cognitively demanding instruction as 1 indicator of high-quality practice. For example, Head Start Program Performance Standards call for curricula that are "sufficiently content rich" and "have an organized developmental scope and sequence that include plans and materials for learning experiences based on developmental progressions" (Head Start Program Performance Standards, 2016). Relatedly, the recent revision to the professional standards and competencies for early childhood educators created by the National Association for the Education of Young Children (NAEYC) articulates that educators need to not only "ask good questions and encourag[e] young children to express and test their own ideas" but also have both content and pedagogical knowledge that can be applied "to integrated curriculum that makes connections through play" (p. 23) (NAEYC, 2019). And, large-scale collection of observational data as part of Quality Rating Improvement Systems (QRIS) has highlighted the need to support cognitively demanding instructional practices that stimulate children's thinking (Pianta et al., 2016).

Yet, most early childhood education programs in the United States do not implement the type of domain-specific, play-based curricula that the classrooms in the current study used (Weiland et al., 2018). Dominating the preschool landscape are whole-child or global curricula that purport to address all domains of child development but whose learning activities do not follow a specified scope and sequence and do not allow for much depth of focus on any 1 domain (Jenkins et al., 2018; Weiland et al., 2018). As reported in our prior work (Authors, 2020), we observed children in this study participating in play-based learning centers in all of the classrooms and that was by far the curricular component that children spent the most time in. As such, there is evidence that children were spending an ample amount of time in activity settings to support the play-based learning dictated by the curriculum and also theorized to support content-rich and cognitively demanding instruction. Further work to measure the constructs of content-rich instruction and cognitive demands across a broader range of programs—and to test whether and how they vary across PreK curricula and different activity settings—is needed to understand how curricula may support these indicators of classroom quality prioritized by existing early learning standards.

Importantly, our exploratory and descriptive analyses also suggested implications for racial equity. We found that classrooms with higher proportions of Black children had lower scores on cognitive demand, on average. This result may in part reflect these classrooms also scoring lower on quality of curricular implementation as well as general emotional support and classroom organization. Although our results are only correlational, it is possible that they may reflect prior research finding that elementary school teachers tend to have lower expectations for Black students, regardless of children's skill levels (Gershenson & Papageorge, 2018). Curricula that aim to support teachers to use cognitively demanding practices may be implemented less well—due to implicit bias and other reasons—in classrooms that include a higher proportion of Black students (Peterson, Rubie-Davies, Osborne & Sibley, 2016). When adopting evidence-based, domain-specific curricula in PreK settings, it may be beneficial for schools and centers to implement targeted strategies to ensure equitable levels of implementation quality and instructional practices across settings.

Content-rich instruction and gains in mathematics skills in prek

Finding that domain-specific curricula can support content-rich instruction is particularly important because results from our predictive models showed that this indicator of classroom quality consistently predicted gains in children's mathematics skills—as

assessed using the WJAP—across the PreK year, over and above a widely-used, existing measure of classroom process quality. In other words, within a sample of classrooms all implementing curricula that aimed to support teachers to expose children to rich content, we found that teachers who were able to do so at higher levels had students who experienced larger gains in mathematics skills during the year than those who implemented these practices at lower levels. This standardized association translates into about 2.5 additional months of learning in mathematics (Hill et al., 2008), which may be particularly important in early childhood contexts given descriptive studies identifying mathematics skills at kindergarten entry—perhaps more so than language/literacy skills—as a critical predictor of academic outcomes through third grade (Duncan et al., 2007).

Although this study is novel in assessing content-rich instruction using a global observational measure, there is growing evidence in the field that content-rich instruction in general supports young students' learning. For example, a recent article by Cabell and Hwang (2020) reviewed the evidence on content-rich instruction in kindergarten to second grade, reporting that the delivery of content appeared to be an effective mechanism for supporting children's language skills and content knowledge, in turn supporting longer-term linguistic and reading comprehension skills. Yet, the bulk of work on content-rich instruction focuses on language and literacy and typically tests the effects of curricula designed to deliver content relative to business-as-usual instruction. In contrast, we examined these practices in a set of classrooms all implementing content-rich curricula and found links between the construct and gains in *mathematics*—but not language—skills. Mathematics instruction integrated with content—for example, talking about different objects in a house and the kinds of shapes they represent while reading a page during a read aloud or discussing strategies for counting the number of seeds during a science activity—may be a mechanism for supporting mathematics in a similar way to how content has been shown to support language skills.

In another example that comes from the mathematics curriculum BPS uses, children are given a sheet of paper displaying a generic outdoor background scene (e.g., hills, stream). They are then given varied manipulatives—like small dinosaur figures or different types of plants—and encouraged to tell a story that relates to topics from science such as herbivores, carnivores, and fossils. They are provided with scaffolds to integrate mathematics into this story, for example by counting the dinosaurs and plants or sorting them by size and shape. In this way, children are exposed to rich and advanced science content while also being explicitly supported in developing mathematics skills. Although we are unable to examine how and whether content-rich instruction varied in the context of mathematics and language/literacy instruction, it is possible that this construct was more salient for supporting mathematics skills because there was more variation in whether and how teachers engaged in content-rich instruction during mathematics. In addition, content-rich instruction is by definition linked to background and world knowledge. In the *Focus on K1* curricula, this may manifest itself as content linked to science, technology, and engineering, which have been associated with children's early mathematics skills (Whittaker et al., 2020).

Importantly, we identified associations between content-rich instruction and gains in children's mathematics skills, over and above the 3 domains of the CLASS and, in our sensitivity tests, over and above overall indicators of intervention fidelity (dosage, adherence, and quality). The CLASS is the most widely used observational measure of classroom quality currently in existence and is used by policymakers and practitioners at federal, state, and local levels to guide program improvement. Recent estimates from Teachstone suggest that about 200,000 to 250,000 CLASS observations occur each year. This measure primarily focuses on interactions between teachers and children, and although it has indicators that aim to capture cognitively demanding instructional practices, its dimensions do not explicitly focus on cognitive demand alone and do not consider content-rich instruction. Recent evidence suggests that the CLASS may not be as predictive of gains in children's skills as prior studies from 10 or more years ago (e.g., Guerrero-Rosada et al., 2021). Work to measure instructional content is important for implementing supports to improve these aspects of quality that may be more predictive of gains in contemporary settings.

Benefits of content-rich and cognitively demanding instruction for children beginning prek with stronger skills

We found that associations between our constructs of interest and children's skills varied depending on the skills children had when they started PreK. Although statistically significant associations varied across outcomes, we generally found that the benefits of content-rich instruction and cognitive demand were limited to children who began the PreK year with *higher* levels of academic skills. Associations between content-rich instruction and children's spring mathematics skills—assessed on the REMA—were larger for children who began PreK with higher mathematics skills, and associations between cognitive demand and gains in children's language skills—assessed using the PPVT—were larger for children who began PreK with higher language skills. Children who began PreK with lower mathematics skills actually had *stronger* mathematics skills in spring—assessed using the REMA—when exposed to *lower* levels of cognitively demanding practices.

The findings align with theory developed through Vygotsky's zone of proximal development (1978), hypothesizing that children who start PreK with higher levels of skills have a stronger knowledge base that allows them to take advantage of and learn from content-rich and cognitively demanding activities because their skill level is more closely matched to the demands of their learning environment. In contrast, children with lower initial skill levels may find some cognitively demanding activities beyond their abilities, even with help from a teacher. There is empirical evidence from other work finding a similar pattern of results. For example, a recent evaluation of the Building Blocks mathematics curriculum—the same cognitively demanding mathematics curriculum used in BPS—found that impacts on mathematics skills were larger for children who began PreK with stronger cognitive skills (Morris, Mattera & Maier, 2016). Although this study did not examine the construct of content-rich instruction specifically, added benefits for higher-skilled children experiencing rich content may operate similarly. Importantly, taken as a whole, children in this study who started the PreK year with weaker skills did make larger gains in academic skills than children who started the year with stronger skills. As such, there continues to be a need to identify the active ingredients driving this differential growth in academic skills for children who start the year with stronger vs weaker skills.

Limitations and directions for future research

This study used a descriptive design, and the analyses do not allow for causal inference. Future experimental research is needed. Second, in our models predicting the REMA mathematics assessment as an outcome, we had to use the WJAP in the fall as a covariate because we were unable to collect the REMA in fall. It is possible that our models predicting the REMA masked associations of interest, an area that future research can address when fitting residualized gains models. Third, we were only able to examine the psychometric properties of the constructs within the current

sample. It is unclear if this measure would demonstrate reliability and be as predictive of children's gains when used in other settings, such as classrooms using one of the whole-child curricula (e.g., Creative Curriculum, HighScope) used more widely across the country. Next, we may have had limited power to detect interaction effects in this paper. A future study leveraging a larger sample of classrooms and students might be better powered to detect small associations between key predictors of interest—content-rich instruction and cognitive demand—and children's skills and to explore heterogeneity in association by children's race/ethnicity, SES, and DLL status.

Fifth, because this measure is a global set of items rated at the end of a classroom observation, we are unable to distinguish *when* content-rich and cognitively demanding instruction occurred, particularly whether it was occurring during activities focused on a specific skill domain. Future work that can disentangle when and how content-rich instruction is delivered can help explain the mechanisms behind the associations found here and inform how to support teachers to provide content-rich instruction. Relatedly, because these global items were completed at the end of an observation where a fidelity tool was being used, we do not know if the items can be used independently. Future research is needed to examine the conditions under which these items can be used and still demonstrate reliability. Finally, we did not include all possible fidelity constructs (Hulleman & Cordray, 2009); child engagement in particular would have added to the richness of our study (e.g., Arbour et al., 2016). We chose only to focus on academic skills as outcomes and did not include children's behaviors, literacy skills, and social-emotional skills in predictive analyses. More research focused on these outcome domains is needed.

Implications

Although this study is exploratory, there are some important implications for future research and practice in this area. First, there is a hunger for feasible and reliable measures of PreK classroom quality that are consistently predictive of child gains. We found that a global observational measure predicted gains in children's mathematics skills, as measured on the WJAP. Although more work is needed to understand why this observational measure did not predict gains in language skills, findings point to the importance of measuring content-rich instruction—or the extent to which teachers deliver background and world knowledge as the medium to support children's skill development—as an aspect of classroom instructional quality. Findings support the need for significant research devoted to better understanding and measuring the key levers in children's PreK learning environments that support their development. Doing so will help the field understand core components of early interventions, such as curricula that deliver rich content (Neuman et al., 2014) and move the needle on children's outcomes. Building observational systems than can enhance quality and outcomes for students, especially those that support high-quality learning for children from historically marginalized and underserved groups, is of paramount importance.

Acknowlegements

The research reported here was conducted as a part of a study funded by Arnold Ventures and Grant R305N160018 – 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 or the U.S. Department of Education. Thanks to the Boston Public Schools, Brian Gold, Blaire Horner, Annie Taylor, the BPS Department of Early Childhood coaches and staff, the BPS Department of Research, the MDRC team (Mirjana Pralica, Rama Hagos, Jennifer Yeaton, Kelly Terlizzi, Desiree Alderson, Marissa Strassberger and Sharon Huang), the Harvard Graduate School of Education research team (Sibyl Holland, Maia Gokhale, and field-based data collection staff), and the University of Michigan research team (Amanda Weissman, Lillie Moffett, Paola Guerrero Rosada, Kehui Zhang). Special thanks to the students and teachers who participated directly in the research.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ecresq.2021.12.010.

References

Arbour, M., Yoshikawa, H., Willett, J., Weiland, C., Snow, C., Mendive, S., … Treviño, E., et al. (2016). Experimental impacts of a preschool intervention in Chile on children's language outcomes. *Journal of Research on Educational Effectiveness, 9*(sup1), 117–149.

Auger, A., Farkas, G., Burchinal, M. R., Duncan, G. J., & Vandell, D. L. (2014). Preschool center care quality effects on academic achievement: An instrumental variables analysis. *Developmental Psychology, 50*(12), 2559.

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

Beck, I. L., & McKeown, M. G. (2007). Increasing young low-income children's oral vocabulary repertoires through rich instruction. *The Elementary School Journal, 107*(3), 251–271.

Bilbrey, C., Vorhaus, E., Farran, D. C., & Shufelt, S. (2010). *Teacher observation in preschool: Tools of the mind adaptation*. Nashville, TN: Peabody Research Institute.

Bloom, H. S., & Weiland, C. (2015). Quantifying variation in Head Start effects on young children's cognitive and social-emotional skills using data from the national Head Start Impact Study. *SSRN Electronic J*.

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 (p. 203–224).

Bronfenbrenner, U., Morris, P. A., Damon, W., & Lerner, R. M. (2006). *Handbook of child psychology.* the ecology of developmental process. Wiley Publishers.

Burchinal, M. (2018). Measuring early care and education quality. *Child Development Perspectives, 12*(1), 3–9.

Cabell, S. Q., & Hwang, H. (2020). Building content knowledge to boost comprehension in the primary grades. *Reading Research Quarterly, 55*(S1), S99–S107.

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.

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

Collins, M. F. (2016). Supporting inferential thinking in preschoolers: Effects of discussion on children's story comprehension. *Early Education and Development, 27*(7), 932–956.

Connor, C. M., Morrison, F. J., Fishman, B. J., Ponitz, C. C., Glasney, S., Underwood, P. S., Piasta, S. B., Crowe, E. C., & Schatschneider, C. (2009). The isi classroom observation system: Examining the literacy instruction provided to individual students. *Educational Researcher, 38*(2), 85–99.

Cook, G. A., Roggman, L. A., & Boyce, L. K. (2011). Fathers' and mothers' cognitive stimulation in early play with toddlers: Predictors of 5th grade reading and math. *Family Science, 2*(2), 131–145.

DeBaryshe, B. D., Gorecki, D. M., & Mishima-Young, L. N. (2009). Differentiated instruction to support high-risk preschool learners. *NHSA dialog, 12*(3), 227–244.

Dickinson, D. K., & Porche, M. V. (2011). Relation between language experiences in preschool classrooms and children's kindergarten and fourth-grade language and reading abilities. *Child Development, 82*(3), 870–886.

Dieterich, S. E., Assel, M. A., Swank, P., Smith, K. E., & Landry, S. H. (2006). The impact of early maternal verbal scaffolding and child language abilities on later decoding and reading comprehension skills. *Journal of School Psychology, 43*(6), 481–494.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., & Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43*(6), 1428.

Duncan, S. E., & DeAvila, E. A. (1998). *PreLAS 2000 technical report*. Monterey, CA: CTB/McGraw-Hill.

Dunn, L. M., & Dunn, D. M. (2007). *PPVT-4: Peabody picture vocabulary test*. New York: Pearson Assessments.

Early, D. M., Maxwell, K. L., Burchinal, M., Alva, S., Bender, R., Bryant, D., … Zill, N. (2007). Teacher education, classroom quality, and young children's academic skills: Results from seven studies of preschool programs. *Child Development, 78*, 558–580.

Farran, D. C., & Bilbrey, C. (2004). Narrative record. In Dale C. Farran (Ed.), *Unpublished instrument available from*. Nashville, TN: Peabody Research Institute, Vanderbilt University.

Fuhs, M. W., Farran, D. C., & Nesbitt, K. T. (2013). Preschool classroom processes as predictors of children's cognitive self-regulation. *School Psychology Quarterly, 28*(4), 347–359.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Gershenson, S., & Papageorge, N. (2018). The power of teacher expectations: How racial bias hinders student attainment. *Education Next, 18*(1), 64–71.

Guerrero-Rosada, P., Weiland, C., McCormick, M., Hsueh, J., Sachs, J., Snow, C., & Maier, M. (2021). Null relations between CLASS scores and gains in children's language, math, and executive function skills. *Early Childhood Research Quarterly, 54*, 1–12.

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.

Hiebert, E. H., Goodwin, A. P., & Cervetti, G. N. (2018). Core vocabulary: Its morphological content and presence in exemplar texts. *Reading Research Quarterly, 53*(1), 29–49.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*(3), 172–177.

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.

Hulleman, C. S., & Cordray, D. S. (2009). Moving from lab to field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness, 2*(1), 88–110.

Institute of Medicine and National Research Council.. (2015). *Transforming the workforce for children birth through age 8: A unifying foundation*. Washington, DC: The National Academies Press.

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.

Morris, P., Mattera, S., & Maier, M. (2016). Making pre-K count: Improving math instruction in New York City. Report produced by MDRC. New York, NY.

Kabay, S., Weiland, C., & Yoshikawa, H. (2020). Costs of the Boston Public prekindergarten program. *Journal of Research on Educational Effectiveness, 13*(4), 574–600.

Maier, M. F., Hsueh, J., & McCormick, M. (2020). *Rethinking classroom quality: What we know and what ae are learning*. New York, NY: MDRC.

McCormick, M. P., Pralica, M., Guerrero-Rosada, P., Weiland, C., Hsueh, J., Condliffe, B., … Snow, C. (2021). Can center-based care reduce summer slowdown prior to kindergarten? Exploring variation by family income, race/ethnicity, and dual language learner status. *American Educational Research Journal, 58*(2), 420–455.

McCormick, M. P., Weiland, C., Hsueh, J., Maier, M, Hagos, R., Snow, C., … Schick, L. (2020). Promoting content-enriched alignment across the early grades: A study of policies & practices in the Boston Public Schools. *Early Childhood Research Quarterly, 52*, 57–73.

NAEYC. (2019). *Early Learning Program Accreditation Standards and Assessment Items*. Retrieved March 9 2021, from https://www.naeyc.org/sites/default/files/globally-shared/downloads/PDFs/accreditation/early-learning/standards_assessment_2019.pdf .

Neuman, S. B. (2006). The knowledge gap: Implications for early education. *Handbook of Early Literacy Rresearch, 2*, 29–40.

Neuman, S. B., Kaefer, T., & Pinkham, A. (2014). Building background knowledge. *The Reading Teacher, 68*(2), 145–148.

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 2332858418784283.

Peterson, E. R., Rubie-Davies, C., Osborne, D., & Sibley, C. (2016). Teachers' explicit expectations and implicit prejudiced attitudes to educational achievement: Relations with student achievement and the ethnic achievement gap. *Learning and Instruction, 42*, 123–140.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system™: Manual K-3*. New York: Paul H Brookes.

Pianta, R., Downer, J., & Hamre, B. (2016). Quality in early education classrooms: definitions, gaps, and systems. *The Future of Children, 26*(2), 119–137.

Purpura, D. J., Napoli, A. R., & King, Y. (2019). Development of mathematical language in preschool and its role in learning numeracy skills. In *Cognitive foundations for improving mathematical learning* (pp. 175–193). New York: Academic Press.

Recht, D. R., & Leslie, L. (1988). Effect of prior knowledge on good and poor readers' memory of text. *Journal of Educational Psychology, 80*(1), 16–20.

Reese, E., & Cox, A. (1999). Quality of adult book reading affects children's emergent literacy. *Developmental Psychology, 35*(1), 20–28.

Ritchie, S., Howes, C., Kraft-Sayre, M., & Weiser, B. (2001). *Emerging academics snapshot*. unpublished measure. Los Angeles: University of California at Los Angeles.

Rojas, N. M., Morris, P., & Balaraman, A. (2020). Finding rigor within a large-scale expansion of preschool to test impacts of a professional development program. *AERA Open, 6*(4), 1–21.

Sarama, J., Clements, D. H., Starkey, P., Klein, A., & Wakeley, A. (2008). Scaling up the implementation of a pre-kindergarten mathematics curriculum: Teaching for understanding with trajectories and technologies. *Journal of Research on Educational Effectiveness, 1*(2), 89–119.

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

Snijders, T. A., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. New York: Sage.

Tseng, V., & Seidman, E. (2007). A systems framework for understanding social settings. *American Journal of Community Psychology, 39*(3–4), 217–228.

Vygotsky, L. (1978). Interaction between learning and development. *Readings on the Development of Children, 23*(3), 34–41.

Wasik, B. A., & Bond, M. A. (2001). Beyond the pages of a book: Interactive book reading and language development in preschool classrooms. *Journal of Educational Psychology, 93*(2), 243.

Wayne, A., DiCarlo, C. F., Burts, D. C., & Benedict, J. (2007). Increasing the literacy behaviors of preschool children through environmental modification and teacher mediation. *Journal of Research in Childhood Education, 22*(1), 5–16.

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.

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.

Weiland, C., & Yoshikawa, H. (2013). Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *Child Development, 84*(6), 2112–2130.

Whittaker, J. V., Kinzie, M. B., Vitiello, V., DeCoster, J., Mulcahy, C., & Barton, E. A. (2020). Impacts of an early childhood mathematics and science intervention on teaching practices and child outcomes. *Journal of Research on Educational Effectiveness, 13*(2), 177–212.

Whorrall, J., & Cabell, S. Q. (2016). Supporting children's oral language development in the preschool classroom. *Early Childhood Education Journal, 44*(4), 335–341.

Woodcock, R. W., Mather, N., McGrew, K. S., & Wendling, B. J. (2001). *Woodcock-Johnson iii tests of cognitive abilities* (pp. 371–401). Itasca, IL: Riverside Publishing Company.

Yudron, M., Weiland, C., & Sachs, J. (2016). *BPS K1DS: Piloting the boston public schools' prekindergarten model in community-based organizations*. Boston, MA: Boston Public Schools.

Zwiep, S., & Benken, B. M. (2012). Exploring teachers' knowledge and perceptions across mathematics and science through content-rich learning experiences in a professional development setting. *International Journal of Science and Mathematics Education, 11*(2), 299–324.