Early Childhood Research Quarterly 63 (2023) 313–326

Contents lists available at ScienceDirect

Early Childhood Research Quarterly

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

Learning experiences vary across young children in the same classroom: evidence from the individualizing student instruction measure in the Boston Public Schools

Christina Weiland[a,*], Lillie Moffett[a], Paola Guerrero Rosada[a], Amanda Weissman[a], Kehui Zhang[a], Michelle Maier[b], Catherine Snow[c], Meghan McCormick[b], JoAnn Hsueh[b], Jason Sachs[d]

[a] *University of Michigan, 610 E University Ave, MI Ann Arbor, Michigan 48104, United States*
[b] *MDRC, United States*
[c] *Harvard Graduate School of Education, United States*
[d] *Boston Public Schools, United States*

ARTICLE INFO

Article history:
Received 7 December 2021
Revised 10 November 2022
Accepted 28 November 2022

Keywords:
Prekindergarten
Kindergarten
Instructional experiences
Child-level measurement

ABSTRACT

Classroom-level quality measures are widely used in early education settings but may mask important variation in learning experiences across children in the same classroom. This study investigates this possibility using detailed data from an observational measure of individual children's learning experiences – Individualizing Student Instruction (ISI). We also examine two other suggested directions for improving early childhood measurement – measuring specific content and learning formats. Our sample includes 263 prekindergarteners and 390 kindergarteners (*M* age=5.2; 51% female; 20% Asian; 20% Black; 32% Latino; 24% White; 4% Other). We found that learning experiences differed substantially across young children enrolled in the same classroom and across student subgroups, particularly for some learning content areas and learning formats. However, this variation did not consistently predict children's language, literacy, math, or executive function gains. The exception was a small relation between time off-task and math gains in both grades, though these findings are sensitive to which math measure is used. Findings underscore the need for more measurement work in early education settings, including development and validation of new instruments and rigorous psychometric studies of existing measures.

1. Introduction

Classroom-level observational measures are widely used to understand the contribution of young children's classroom learning experiences to their development. Such measures typically provide classroom-level estimates of the quality of routines, the environment, and interactions between teachers and students (e.g., Pianta et al., 2008; Harms et al., 1998). These measures have strengths, such as facilitating cross-system comparisons of programs and contributing to large-scale quality improvement efforts (Chaudry et al., 2017; Bassok et al., 2019; U.S. Department of Health and Human Services, 2018). However, they do not consistently predict gains in preschoolers' and kindergarteners' developmental outcomes (Guerrero Rosada et al., 2021; Weiland et al., 2013; Zaslow et al., 2016). Accordingly, scholars have called for a next generation of measurement work in early childhood education (Burchinal, 2018; Weiland, 2018). One suggested direction is focusing on children's individual learning experiences with the idea – described in this paper as the "masking hypothesis" – that classroom-level observational measures may mask variation across children in the same classroom. Another is measuring specific instructional content and format of instructional activities.

To help meet this call from scholars for a next generation of measurement work, we use data from what, to our knowledge, is the most detailed observational measure of individual children's classroom learning opportunities and experiences developed to date – the Individualizing Student Instruction measure (ISI; Connor et al., 2009). The ISI measures three sources of within-classroom variation in children's experiences: the amount of time a child is engaged in learning activities (versus in management/routines or off- task); the amount of time a child is exposed to different content areas (literacy, math, science, etc.); and the amount of time the child spends in different learning formats (whole group, small group, individual, etc.). In this video-tape-

[*] Corresponding author.
E-mail address: weilandc@umich.edu (C. Weiland).

https://doi.org/10.1016/j.ecresq.2022.11.008
0885-2006/© 2022 Elsevier Inc. All rights reserved.

based coding system, children are continuously followed during the observational period, with highly specific codes for the content of their learning.

Using data from the ISI and a sample of children enrolled in Boston Public Schools (BPS) prekindergarten and kindergarten classrooms in 2016–2017 and 2017–2018, respectively,[1] we focus on three aims: how learning experiences vary across children in the same classroom, including across content areas and learning formats; whether student characteristics (dual language status, gender, free-reduced-lunch status, race/ethnicity, and baseline skills) predict their learning experiences; and whether students' learning experiences predict gains in their language, literacy, mathematics, and executive function (EF) skills in prekindergarten and kindergarten.

Our study adds to the literature in several ways. First, the ISI is fine-grained, and we enhanced it further in order to examine nuances in the instructional experiences offered to individual children as thoroughly as we could. For example, we expanded on specific instructional codes for children's math experiences (Marks et al., 2016), and are the first to report math observation results using the ISI. Furthermore, because the ISI does not measure quality of instruction, we also coded study classrooms using the widely used Classroom Assessment Scoring System (CLASS; Pianta et al., 2008). This first-ever use of the CLASS and the ISI together allowed us to distinguish individual children's learning experiences from the overall quality of instruction in predicting their learning gains. Finally, our study sample is demographically diverse with respect to language, family income, and race/ethnicity, enhancing the applicability of our results to other settings and allowing a direct test of how learning experiences may vary across important student subgroups. This latter question is new to ISI studies and represents a potential key advantage of child-level over classroom-level measures. By definition, child-level measures can potentially detect differences in learning experiences across subgroups while classroom-level measures cannot. Capturing child-level information on instructional experiences may thus be important for increasing equity and improving children's outcomes in the early years.

2. Theoretical framework

Our study draws from several different complementary theories about how instructional content and learning formats, particularly when measured at the child level, may drive differential gains in early learning. For content and learning formats, a new conceptual framework from Maier & colleagues (2020) separates classroom interactional experiences (e.g., teacher emotional support) from instructional experiences such as the content and activity learning format (i.e., whole group, small group, centers). The authors refer to these components as the "how" and the "what" of classroom instruction. We further differentiate language, literacy, and math content into *constrained* versus *unconstrained* skills, following McCormick & colleagues (2022), Snow & Matthews (2016), and Paris (2005). "Constrained" refers to skills for which there is a finite ceiling, like knowing the names and sounds of letters. "Unconstrained" refers to skills that children build throughout their lifespans, like vocabulary and problem solving. Both are important to children's literacy and mathematics development and to school success. In the early years, students may receive relatively more constrained skill instruction and relatively less unconstrained instruction, making measuring students' experiences of these two types of instruction important for understanding their within-year gains and gains across time.

Transactional developmental theory (Sameroff, 2009) helps to articulate *why* individual young children may differentially experience content and learning formats within the same classroom. That is, transactional developmental theory posits that teacher behaviors—or the learning experiences they may offer to children—interact with child skills and environmental inputs, which can then have a reciprocal effect on children's behavior (Sameroff, 2009). For example, a child who always chooses the block area during center time might experience more math-centric interactions with her teacher than a peer who prefers the book corner. This is particularly relevant in prekindergarten and kindergarten programs like Boston's that emphasize child choice and thus where children's preferences, temperament, and interests drive some aspects of their learning experiences.

Vygotsky's zones of proximal development add a related, more specific lens on teacher's decisions and behaviors with individual young learners. This theory emphasizes what a learner can do independently versus with scaffolding supports (Vygotsky, 1978). Following this theory, beyond the child's preferences and choices, a teacher might offer different learning opportunities in preschool and kindergarten to a child with an advanced vocabulary but weak math skills versus a peer with the opposite profile.

Finally, teachers' own implicit and explicit beliefs and biases may influence young children's individual learning experiences (Alvidrez & Weinstein, 1999). For example, early educators may respond more negatively to Black students, particularly boys, than to their White peers (Gilliam et al., 2016), and perceive young boys as more mathematically competent than girls (Robinson-Cimpian et al., 2014). These beliefs and biases can lead classmates with different demographic characteristics to have different relationships with the same teacher and different classroom learning experiences.

Importantly, the most widely used observational measures in the field do not measure either content or learning formats, nor do they measure children's individual learning experiences (Weiland & Guerrero Rosada, 2022), limiting the ability to test these complementary theories empirically in many existing data sets. Our study with the ISI helps to address this gap.

3. Previous research on measures of individual children's learning experiences, instructional content, and learning formats

Existing research provides some empirical evidence that observational measures can detect meaningful variation in young children's individual learning experiences, instructional content, and learning formats. For example, work to date with the ISI shows that it is sensitive to detecting differences in individual children's classroom experiences (Connor et al., 2010; Day et al., 2015). In preschool specifically, research with the ISI has found substantial variability in individual children's time spent in meaning-focused, individual literacy activities (M= 11.62 min, SD= 15.00, range 0.00–66.20) and code-focused, teacher-led literacy skills (M=.89 min, SD=6.11, range 0–65.02; Connor et al., 2006).

Child-level measures can also identify differences in learning experiences by children's background characteristics, an important equity concern that classroom-level measures cannot assess. Across several studies, the Emerging Academic Snapshot (the SNAP) has identified differences by children's family income, race/ethnicity, and gender in time spent in free choice and teacher-directed activities as well as literacy and math activities (Early et al., 2010; Pianta et al., 2005). However, another child-level measure – the Language Interaction Snapshot (LISn) – found no evidence of dif-

[1] The Boston Public Schools refers to its public preschool program for four-year olds as "prekindergarten." When describing the Boston program in this paper, we similarly use the term "prekindergarten." However, in later sections of the paper when discussing the broader literature on early childhood care and education, we use the term "preschool."

ferences in language learning experiences by child dual language learner (DLL) status (Bratsch-Hines et al., 2019), either because there were no such differences or due to instrument insensitivity. To date, the ISI has not been used to study differences by child background characteristics. Accordingly, this is a key contribution that our study makes to the literature.

In terms of predictive validity of child-level measures, associations have ranged from null to statistically significant but small, for both content and learning formats (d=.15–.28; Bratsch-Hines et al., 2019; Burchinal et al., 2021; Chien et al., 2010; Howes et al., 2008; Sabol et al., 2018). The ISI follows this same pattern. For example, in a sample of 156 preschoolers, time spent on whole-group or individual activities led by a teacher predicted alphabet and letter-word gains (β =.27), and child-led experiences, such as free play, predicted gains in vocabulary (β=.25; Connor et al., 2006). In a kindergarten ISI study, total amount of time off-task predicted fewer gains in letter-word knowledge (β =-.21), math (β =-.22), and reading comprehension at the end of 1st grade (β =-.25; Moffett & Morrison, 2020). However, time spent in vocabulary and phonics instruction did not predict gains in literacy or language skills in one kindergarten ISI study (Al Otaiba et al., 2008), nor did time spent in literacy activities predict gains in vocabulary or decoding skills during first grade (Connor et al., 2004) or gains in literacy scores during third grade (Connor et al., 2014). More research is needed, particularly in preschool and kindergarten.

Finally, some studies have shown classroom-level observational measures of learning formats and specific content to be sensitive to differences in teacher practices and to predict children's gains. For learning formats, some studies have found that more time in whole-group settings, compared to free choice or child independent learning, predicted child learning gains in language and literacy (Ansari & Purtell, 2017; Fuligni et al., 2012). Other studies have found that more time in free-choice and less time in whole-group in preschool was associated with higher gains in social-emotional skills (Fuligni et al., 2012) math (Chien et al., 2010) and higher ratings of classroom quality (Nores et al., 2022).

For content, available observational measures typically target a specific curricular approach (e.g., curriculum fidelity tool) or subject area, like math or literacy. For example, the Classroom Observation of Early Mathematics Environment and Teaching (COEMET) measure assesses the classroom math culture and quality of math activities by capturing nuance in the content of math instruction (e.g., number sense, geometry etc.), and the teacher strategy used to teach it (e.g., open-ended questioning) (Sarama & Clements, 2009). The quality of math instruction as measured by COEMET has shown some predictive validity in prior studies for preschoolers' math gains and has been shown to partially mediate the effects of the Building Blocks math curriculum (Sarama et al., 2008). In language and literacy, the Early Childhood Language and Literacy Classroom Observation Tool (ELLCO Pre-K) (Smith et al., 2008) and the Observation of Language and Literacy Instruction (OLLI) (Guo et al., 2012) assess nuances in language and literacy-specific classroom activities. These measures, however, have not been very widely used and studies of their predictive validity are mixed.

4. Present study

We use data from the ISI to contribute to a next generation of measurement work in early education (Burchinal, 2018; Weiland, 2018). Specifically, we explore three research questions:

- How do learning experiences, including content and learning formats, vary across children within the same classroom in prekindergarten and kindergarten?
- Do children's individual learning experiences vary across subgroups defined by dual language learner status, gender, eligibility for free-or-reduced-price lunch, race/ethnicity, and baseline skills?
- Do children's individual learning experiences predict gains in their language, literacy, math, and EF skills, over and above classroom-level quality in prekindergarten and kindergarten?

5. Method

5.1. Participants and setting

Our sample consists of 263 prekindergarten students in the 2016–2017 academic year, recruited from 39 classrooms within 19 public elementary schools offering the Boston Public Schools prekindergarten program. We followed our prekindergarten sample into their kindergarten classrooms (2017–2018). We lost 56 students due to attrition and recruited an additional 183 students from the same kindergarten classrooms. The kindergarten sample is comprised of 390 students in 51 kindergarten classrooms within 20 schools.

The prekindergarten program is free, full-day, and open to any age-eligible child for the academic year (though there is more demand than supply; Weiland et al., 2020). About 92% (N = 36) of the prekindergarten classrooms included in the current study used the BPS *Focus on K1* curriculum, an adapted version of Opening the World of Learning (Schickedanz & Dickinson, 2004), 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, and several district-developed components. In kindergarten, 86% (N = 44) of classrooms used the BPS *Focus on K2* curriculum, a district curriculum designed to be vertically aligned with the prekindergarten curriculum (McCormick, Hsueh, Weiland, & Bangser, 2017).

As shown in Table 1, study children were diverse with respect to race/ethnicity, Dual Language Learning status, eligibility for free-or-reduced price lunch, and gender. For example, 30% were Latino, 28% White, 17% were Black and 17% were Asian in the prekindergarten sample, with roughly similar racial/ethnic composition in the kindergarten sample. Overall, 57% of the prekindergarten sample and 61% of the kindergarten sample was free-reduced-lunch eligible. Teachers were highly experienced (e.g., 15 years average teaching experience in our prekindergarten sample and 12 years for kindergarten), racially and ethnically diverse (prekindergarten sample: 9% Asian, 21% Black, 19% Latino, 47% White; kindergarten sample: 6% Asian, 14% Black, 19% Latino, 61% White), and most held master's degrees (82% in both samples).

5.2. Procedures

The Institutional Review Boards at the lead and partner organizations for this study approved the human subjects plan prior to the commencement of study activities. For parsimony, additional details on procedures are in Appendix C.

5.2.1. School and classroom recruitment

In 2016, we randomly selected 25 public schools from the 76 schools in the district offering public prekindergarten. Ultimately, 19 schools and 96% (N = 39) of general education and inclusion-model prekindergarten teachers in these schools agreed to participate in the study and videotaping activities. We followed the participants into kindergarten in the fall of 2017. Students were spread across 26 schools and 54 kindergarten classrooms. All teachers were asked to participate and 95% agreed to participate in the

Table 1
Child demographics and assessments, ISI measurement, and CLASS descriptive statistics.

	Prekindergarten (*n*=263 children)			Kindergarten (*n*=390 children)		
	Mean	SD	Percent Missing	Mean	SD	Percent Missing
Child characteristics						
Race/ethnicity						
Latino	0.30	-	0.00	0.32	-	0.00
White	0.28	-	0.00	0.23	-	0.00
Black	0.17	-	0.00	0.18	-	0.00
Asian	0.17	-	0.00	0.22	-	0.00
Other race	0.08	-	0.00	0.04	-	0.00
Female	0.52	-	0.00	0.50	-	0.00
Eligible for free/reduced-lunch	0.57	-	0.00	0.61	-	0.00
Dual language learner	0.55	-	0.00	0.59	-	0.00
Child age at baseline	4.66	0.29	0.00	5.60	0.30	0.00
Attended CBO	-	-	-	0.03	-	0.00
Parent education						
High school diploma/GED or less	0.31	-	1.52	0.31	-	6.67
Two-year degree or Equivalent	0.23	-	1.52	0.28	-	6.67
Four-year degree	0.16	-	1.52	0.17	-	6.67
Advanced degree	0.30	-	1.52	0.23	-	6.67
Age of mother at first child's birth	27.72	6.93	1.90	26.95	6.82	9.23
Household size	4.29	1.27	2.28	4.29	1.26	6.67
Whether at least one adult in household works full time	0.88	-	1.52	0.88	-	6.67
Married/parent	0.64	-	1.52	0.60	-	6.67
Parent age	36.55	7.57	2.66	32.57	9.21	6.67
Fall achievement measures						
PPVT raw	73.89	28.02	1.52	87.20	28.72	1.79
WAP raw	12.72	5.04	1.52	16.00	5.27	1.02
Digit span	3.13	1.06	1.52	3.50	0.99	1.28
REMA raw	-	-	-	11.36	5.70	2.56
DIBELS FSF				15.59	12.86	22.56
DIBELS LNF				23.84	16.66	22.56
Spring achievement measures						
PPVT raw	87.41	26.86	1.52	101.41	27.43	1.53
WAP raw	15.95	4.49	1.90	19.51	4.58	1.28
Digit span	3.51	1.02	1.52	3.80	0.81	1.53
REMA raw	17.37	8.72	1.52	16.48	7.96	0.01
DIBELS LNF	-	-	-	50.82	17.85	17.94
DIBELS PSF	-	-	-	40.77	18.65	17.94
DIBELS NWF WWR	-	-	-	7.19	11.03	17.94
ISI measures (child level)						
Minutes observed	184.37	63.88	0.00	218.91	59.62	0.00
One observation	0.16	-	0.00	0.15	-	0.00
Classroom quality (classroom level)						
Classroom organization	5.47	0.60	0.00	5.83	0.61	0.00
Instructional support	3.26	0.64	0.00	2.49	0.60	0.00
Emotional support	5.62	0.61	0.00	5.75	0.49	0.00

broader study. Due to funding constraints, we were unable to film all classrooms. We prioritized classrooms with the most student participants, resulting in a sample of 51 kindergarten classrooms and 20 schools.

5.2.2. Student recruitment

In prekindergarten fall, 81% of children in participating classrooms agreed to participate. We randomly selected 50% (~6–10 per classroom) of consented children to participate in student-level data collection activities, for a total of 307 prekindergarten students. We were then able to collect videotapes of prekindergarten classroom experiences for 263 (86%) of these students. In kindergarten, we consented and recruited 78% of kindergarten students in the participating classrooms ($N = 220$) who had not attended the Boston prekindergarten program in the 2016–2017 academic year and thus who had not been consented in the previous year, for a total of 483 consented kindergarten students. Of these consented and selected prekindergarten and kindergarten students, we excluded students who were either not present for filmed observations or whose classrooms were not filmed due to capacity constraints ($N = 44$ prekindergarten students and $N = 93$ kindergarten students). Our final sample size was 263 for prekindergarten and 390 for kindergarten ($N = 207$ from the prekindergarten sample and $N = 183$ students enrolled in kindergarten).

5.2.3. Parent surveys

We collected parental demographic information via 20 min parent surveys in the fall of prekindergarten (fall of 2016) and again in the fall of kindergarten (fall of 2017). Although the majority of parents completed the survey in English, we also translated the survey into Spanish, Vietnamese, and Mandarin. Parents received a $25 gift card for completing the survey. Of the 263 students in the first year of our study, 255 (97%) had parents who completed the survey in at least one of the years. Of the 390 students in the second year of our student, 345 (88%) had parents who completed the survey in at least one of the years.

5.2.4. Direct assessments

Prekindergarten students were assessed in the fall of 2016 (October 1st through December 12th) and the spring of 2017 (April 5th through June 16th). We assessed kindergarten children in the fall of 2017 (October 1st through December 12th) and spring of

2018 (April 5th through June 16th). All child assessors were trained to reliability. A master's-level supervisor observed 10% of field assessments to ensure high-quality administration. Before beginning the study battery, in both prekindergarten and kindergarten, assessors used the Pre-language Assessment Scale (preLAS; Duncan & DeAvila, 1998) to determine the administration language for a subset of assessments. Overall, 14% of the prekindergarten sample completed a subset of assessments in Spanish in fall 2016 and 4% did so in spring 2017. In kindergarten, 4% completed a subset of assessments in Spanish in the fall of 2017 and 1% did so in spring 2018.

5.2.5. Classroom observations

Prior to conducting observations, we reviewed all teachers' weekly schedules and identified a two-to-three hour block of instructional time which included substantial focus on language/literacy and another two-to-three-hour block with some focus on math instruction. Video length was dependent on the teacher's schedule and any other activities happening at the school that day. We observed prekindergarten classrooms across two school days between January 25 and May 10, 2017 (89% of classrooms were observed in February and March). On average, the second observation occurred 13.78 days ($SD = 13.65$) after the first. Classrooms were observed for an average of 3.16 h total ($SD = .83$, range=2.21–4.62 h).

We observed kindergarten classrooms across two school days between January 16 and April 5, 2018 (98% of classrooms were observed in February and March). On average, the second observation occurred 6.47 days ($SD = 5.62$) after the first observation. Classrooms were observed for an average of 3.73 h total ($SD = .64$, range=2.30–4.98 h).

We used two video cameras during each observation session, one focused primarily on the teacher (and the teacher's microphone), and the other on the students. Before coding, we synchronized videos from the two observations to effectively track both the teacher and students as they moved between camera angles. We used Noldus Observer XT 13 software for coding videotapes with the ISI observation measures (Noldus Information Technology, 2013). In addition, we collected relevant classroom information that may not be evident from the videotape (for example, sample worksheets and descriptions of activities that took place outside the camera's view). This information helped answer questions about the content of the activities children were observed doing.

Coders participated in multiple training sessions on ISI measures and were tested on the mastery of the codebook before coding. After training, coders had to show reliability on the ISI via coding four 20 min video segments. Compared to a master-coded file, all coders scored >.80 Kappa on each of the four videos. Throughout the coding process, to prevent drift in reliability, we randomly selected and double-coded 20% of the video observations. After each round of double coding (five total rounds), coders discussed any coding disagreements. We calculated reliability in the Noldus Observer XT software which compares the duration of time (start and end time) of each code and the order/sequence of codes within a 15 s grace window. For prekindergarten, coders' average Kappa was .76 and for kindergarten, .73, a very similar level of reliability as past ISI studies (e.g., average of .76 in Connor et al., 2009).

For the Classroom Assessment Scoring System (CLASS; Pianta et al., 2008), coders participated in a two-day training to learn the CLASS measure and then established reliability on a set of master codes created by the test developers. As recommended by the measure's protocol (Pianta et al., 2008), coders used cycles of 20 min for observing and 10 min for scoring, which they repeated up to 4 times for each observation. Coding began when instruction commenced in the video and ceased after 80 min of observed time. We double-coded 20% of the observations to assess interrater reliability. The final ICCs representing interrater reliability (within 1 point) for the three domains were 96% for Emotional Support, 94% for Classroom Organization, and 88% for Instructional Support. We also conducted drift checks wherein observers had to code a master tape every three weeks to ensure they stayed reliable across time.

5.3. Measures

5.3.1. Children's individual learning experiences

To capture child-level learning experiences, we used the Individualizing Student Instruction (ISI) Coding System (Connor et al., 2009). The ISI is not a quality rating scale measure, but rather captures a continuous measure of *quantity of time* (e.g., 0–58 min) an individual child is engaged in different learning experiences. Specifically, the ISI measures: the amount of time a child is engaged in learning activities (versus in management/routines or off task); the amount of time a child is exposed to different content areas (literacy, math, science, etc.); and the amount of time the child spends in different learning formats (whole group, small group, individual, etc.). Our version of the ISI included math instructional codes and more nuanced non-instructional codes (Marks et al., 2016). We also added in codes that captured whether each math activity was close-ended (one right answer) or open-ended (many right answers)/flexible (one right answer but with multiple approaches). We consulted with a math expert (REDACTED) to define these codes and checked reliability on them in the training and coding processes just as we did on all other ISI codes. For each target child, our coders coded each second of observed time, switching codes as necessary to capture children's focal activities, content, and learning formats.

Appendix A Tables 1 and 2 provide breakouts of our overarching language and literacy and math constructs. Within literacy, for example, we coded the focal content of each activity each child engaged in using codes such as phonological awareness (e.g., child is learning what the letter "b" sounds like), fluency (e.g., child is practicing reading text smoothly), and text reading (e.g., teacher is reading a story book to the child). In math, we did the same, with codes such as number sense and concepts (e.g., child is counting out objects), operations (e.g., child is practicing adding and subtracting one object), and data analysis (e.g., teacher askes child to compare the amounts of objects by creating a graph). Following McCormick & colleagues (2022), Snow & Matthews (2016), and Paris (2005), we grouped language and literacy and math codes into both total codes and into more constrained versus less constrained skills (see Appendix C for more details).

In addition to instructional time, we also coded the nature of children's non-instructional time. For example, we coded when a child was engaged in off-task behavior (e.g., child is supposed to be looking at a book but is instead drawing on the white board) or participating in a routine or exposure to teacher management practices (e.g., teacher gives directions and child cleans up their play area before moving on to the next activity).

For the students with two observations (84% in prekindergarten, 85% in kindergarten, Table 1), we first summed their data across both observation days to create aggregate child-level measures for each ISI construct of interest. We then followed previous ISI studies and used the number of minutes in a specific code at the child-level as our primary analysis strategy (Connor et al., 2006; Day et al., 2015). We used the percentage of time spent in a specific code as a robustness check (described further below). To account for the students who only had one observation, we included a dichotomous indicator as a covariate in our regression models (see analytic section).

5.3.2. Classroom process quality

We measured classroom process quality using the Classroom Assessment Scoring System (CLASS) PreK (Pianta et al., 2008). This observational tool measures three domains of teacher-child interactions: Emotional Support, Classroom Organization, and Instructional Support. All the dimensions are scored on a 7-point scale, with higher scores indicating higher quality. In prior work with our prekindergarten sample, the CLASS did not predict gains in children's outcomes (Guerrero Rosada et al., 2021). To disentangle quantity of instruction (ISI) from quality (CLASS), we controlled for CLASS Instructional Support in our key regression models. See Table 1 for descriptive statistics.

5.3.3. Receptive vocabulary

To capture children's vocabulary, we used the Peabody Picture Vocabulary Test-IV (Dunn & Dunn, 2007), which has been normed and used widely in diverse samples of children in the U.S (Puma et al., 2010). Its test– retest reliability is 0.93, and it has shown qualitative and quantitative validity properties (Dunn & Dunn, 2007). Children are asked to choose (verbally or nonverbally) which of four pictures best represents a stimulus word. Following other studies of four- and five-year-old children (Weiland & Yoshikawa, 2013; Wong et al., 2008), we used the raw score total as our outcome measure. We assessed all children on the PPVT in English regardless of their results on the PreLAS language screener (explained above) in order to obtain an English receptive language score for the full sample.

5.3.4. Literacy

In kindergarten only, teachers administered the Dynamic Indicators of Basic Literacy Skills NEXT (DIBELS; Good et al., 2011). Administered subtests measured children's letter knowledge (Letter Naming Fluency; LNF), phonological awareness (Phoneme Segmentation Fluency, PSF; and First Sound Fluency, FSF), and alphabetic principles (e.g., letter-sound correspondence and the ability to blend letters into words; Nonsense Word Fluency Correct Letter Sounds and Nonsense Word Fluency Whole Words Read; NWF CLS and NWF WWR). Following test developer recommendations, LNF and FSF were administered in the fall. All subtests except for FSF were used during the spring. Because the two NWF subtests were highly correlated in our sample (r=.90, $p < .001$), we used just the NWF WWR subtest, which had a much lower mean than the other DIBELS subtests and thus appears to have been harder for students than other subtests (see Table 1). These subtests have high reliability (.9 or above), are widely used, are sensitive to intervention effects, and have good concurrent, predictive, and discriminant validity properties (Good et al., 2011).

5.3.5. Math skills

To assess early numeracy skills, we used the Woodcock-Johnson Applied Problems subtest (Woodcock et al., 2001). Its estimated test–retest reliability for 2- to 7- year-old children is 0.90 (Woodcock et al., 2001) and it has been nationally normed and used with diverse populations of children (Gormley et al., 2005; Wong et al., 2008). We assessed Spanish-speaking children who did not pass the PreLAS language screener using the equivalent Spanish language version from the Batería III Woodcock Muñoz (Schrank et al., 2005).

The Woodcock-Johnson Applied Problems subtest has been criticized by some math 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, we also used the Research-based Early Math Assessment (REMA; Clements et al., 2008; Weiland et al., 2012), a hands-on, one-on-one assessment of children's early math skills (e.g., numeracy, geometry, operations, spatial reasoning). The REMA includes manipulatives and more items targeted to the early childhood period. The alpha reliabilities of the test subscales range from r = .89 (number) to .71 (geometry; Clements et al., 2008). We used the REMA raw score.

5.3.6. Executive function

We used the Forward Digit Span Assessment (FDS) to measure children's working memory, one of the key components of executive function. The FDS requires that children repeat several series of numbers in rapid succession, with an increasing number of digits presented once the child has successfully repeated a prior sequence (Wechsler, 1974). It measures the phonological loop component of working memory. FDS has high correlations with other EF tasks, and has good test-retest reliability in young children (r =.80; Muller et al., 2012). We used the categorical score for Forward Digit Span, which represents the sequence with the highest number of digits that the child repeated accurately.

Child-level covariates. We used administrative data from the school district to create child-level covariates and subgroup indicators. Following recommended best practice (Gehlbach & Robinson, 2018), we selected covariates to match those in other published papers that use this same dataset (Guerrero Rosada et al., 2021; McCormick et al., 2021; 2022). We used a set of indicators to describe children's race/ethnicity (Black, Latino, Asian, or Other Race/Ethnicity, with White as the reference group). We used dichotomous indicators to capture whether each child was eligible for free-reduced-priced lunch, female, and/or a Dual Language Learner (DLL; determined based on parent's report that a language other than English was spoken at home, was the language most often spoken by the student, or was the student's first language). We also used the child's birth date to calculate age at the time of the Fall 2016 assessment in the prekindergarten models and Fall 2017 assessment in the kindergarten models. Lastly, in the kindergarten models, we controlled for whether students attended preschool in a BPS-affiliated community-based organization (CBO) during prekindergarten year. BPS-affiliated CBOs used a similar curriculum to the public prekindergarten program of interest in our prekindergarten model but differed on features such as teacher education and experience (McCormick et al., 2020). Of the 390 students in our kindergarten sample, only 12 (3%) attended a CBO the year before they entered kindergarten.

5.4. Analytical approach

5.4.1. Missing data

We had overall low rates of missingness (0–1.9%), with no evidence of systematic missingness on study variables. The exception was the teacher-administered DIBELS subscales in kindergarten (17–23% missing). Accordingly, we used complete case analysis but included a robustness check in which we imputed missing data for models that used the DIBELS.

5.4.2. Multilevel modeling

To answer our first research question – how learning experiences vary across children within the same classroom – we fit multilevel models using Eq. (1):

$$ISI_{ijk} = \beta_0 + \beta_1 Time_{ijk} + \beta_2 NumObs_{ijk} + u_k + \gamma_{jk} + e_{ijk} \qquad (1)$$

Where the subscript i refers to an individual student, j denotes an individual classroom, k represents an individual school, ISI is the ISI outcome of interest, $Time$ represents the number of minutes an individual student was observed, $NumObs$ refers to the number of observations per student, u_k and γ_{jk} are the school and classroom residual terms, respectively, and e is the student-level residual term. We then calculated the percent of the variance at each level (child, classroom, school) using a standard intracluster correlation (ICC) to identify whether and to what degree learning experiences as measured by our ISI constructs varied across children in the same classroom.

To answer the second research question – whether student characteristics predict learning opportunities – we added to Eq. (1) an indicator for whether child *i* was a member of a subgroup of interest (i.e., a child was a dual language learner, female, free-reduced-priced-lunch eligible). We followed the same strategy for race/ethnicity with indicators for Black, Asian, Latino, and Other race or multiracial (with White as the reference category) and for whether the student scored in the lowest or highest quartile on the PPVT and Woodcock-Johnson Applied Problems subtests administered in the fall (with the middle two quartiles as the reference category). For kindergarten only, we also used an indicator for whether the child attended Boston prekindergarten, given hypotheses in the literature that teachers may respond differently to prekindergarten attenders (Weiland, et al., 2021). The coefficient on the subgroup variable and its associated standard error are the parameters of interest, identifying whether learning experiences differ by student demographics.

For research question three – whether individual children's learning experiences predict growth in their skills across the school year – we fit residualized change models:

$$\begin{aligned} Child\_outcome_{ijk} = {} & \beta_0 + \beta_1 ISI_{ijk} + \beta_2 Baseline_{ijk} + \beta_3 CLASS_{jk} \\ & + \beta_4 Time_{ijk} + \beta_5 NumObs_{ijk} + \beta_6 NumDays_{ijk} \\ & + X_{\boldsymbol{ijk}}\Gamma + Z_{\boldsymbol{ijk}}\mathrm{P} + u_k + \gamma_{jk} + e_{ijk} \end{aligned} \tag{2}$$

where *Child_outcome* refers to the relevant child-level outcome measure; *Baseline* denotes the relevant fall child-level assessment for the outcome; *CLASS* is CLASS Instructional Support measured at the classroom level; *X* is a vector of student level characteristics (race/ethnicity, female, dual language learner, free-reduced-price-lunch eligible, and child age at baseline); *Z* is a vector of parent covariates measured at the student level (parental education, age of mother at first child's birth, household size, whether at least one adult in household works fulltime, married/partner, and parent age); and *NumDays* denotes the number of days between the baseline and outcome assessment; and all other terms are defined as in Eq. (1). Kindergarten models also included a control for whether kindergarten student *i* attended Boston prekindergarten in the previous school year. All other terms are defined as in Eq. (1). In Eq. (2), β_1 and its associated standard error identify the association between each ISI construct of interest and children's learning gains. We did not correct for multiple comparisons, following the advice of Schochet (2009) for exploratory work like ours.

6. Results

6.1. Descriptive statistics

As shown in Table 2, prekindergarteners spent 122 min (66%) on instruction and kindergarteners, 164 min (74%; see Appendix A Table 3–5 for all number of minutes descriptives in percentages). Prekindergarteners spent the remainder of the observed time in management/routines (51 min, or 28%) and off-task behavior (20 min, or 11%). Kindergarteners spent the remainder of their time in management/routines (50 min, or 23%) and off-task behavior (15 min, or 7%). Both groups spent the majority of time in non-whole group settings across small group, centers, and individual learning formats (105 min, or 57%, in prekindergarten; 118 min, or 54%, in kindergarten) and the rest in whole group.

In terms of content of instruction, children spent the most time in language and literacy (65 min or 35% of all observed time for prekindergarteners; 124 min or 57% for kindergarteners). As shown in Table 2, within language and literacy, prekindergarten children spent 16 min on constrained skills and 35 min on unconstrained skills. Kindergarten children spent 38 min on constrained skills and 41 min on unconstrained skills.

Children spent the next-highest amount of instructional time in math (36 min, or 20%, prekindergarten and 43 min, or 20%, in kindergarten). Both grades spent far more time on constrained than unconstrained math skills (21 min versus 2 min in prekindergarten and 20 min versus 4 min in kindergarten; see Table 2). Prekindergarteners spent slightly more time on closed versus open and flexible-ended math activities (12 versus 11 min, respectively). Kindergarteners spent more time on open and flexible-ended activities than close-ended (15 min versus 8 min, respectively).

At the classroom level, correlations between our key ISI constructs and the CLASS were mostly small and not statistically significant (see Appendix A Table 6). The percent of time children were off task was most consistently and strongly correlated with the three CLASS domains (r=-.44 to -.51 in prekindergarten and r=-.22 to -.45 in kindergarten). Unconstrained language and literacy also showed small, consistent, and statistically significant correlations with all three CLASS domains, in kindergarten only (r=.36 to .38). Underscoring that the ISI and CLASS measure different constructs, total instruction on the ISI showed a small negative, statistically non-significant correlation with CLASS Instructional Support in prekindergarten (r = -.04, p > .10) and a small positive correlation in kindergarten (r = .20, p > .10).

7. Variation in learning experiences across children in the same classroom (RQ 1)

We found that the ISI is sensitive to detecting variation in learning experiences across children in the same classroom in both prekindergarten and kindergarten (Table 3). There was variance at the child level for all ISI constructs (ICC range of 3%–74%). Overall, we found more variance in ISI constructs at the child level in prekindergarten than kindergarten.

For content of instruction, most of the variance in unconstrained (74%) language and literacy skills was at the child level in prekindergarten, while in kindergarten, it was evenly split with 48% at the child level and 52% at the teacher level. For constrained skills, for both grades, the teacher level explained most of the variance. For math, in prekindergarten, the pattern was similar to the language and literacy findings: variation in total instruction and constrained skills was mostly at the teacher level (54%–70%), while most of the variation in unconstrained skills was at the child level (63%). In contrast, most of the variation in all math ISI constructs in kindergarten was at the teacher level (range 51%–71%). Finally, more of the variation in time on instruction and in off-task behavior was at the child level in prekindergarten than in kindergarten, when teachers or schools accounted for more of the variation. In terms of format, teachers drove variation in both prekindergarten and kindergarten (range of 73%–92%).

8. Student characteristics as drivers of learning experiences (RQ 2)

As shown in Table 4, we found that the ISI is sensitive to detecting differences in children's learning experiences by their demographic characteristics in both grades. In prekindergarten, girls spent about 9 min more on instruction than boys (d=.18, p < .0001), with most of that extra time allocated to literacy (7 more minutes than boys; d=.25, p < .0001) and within literacy, about 5 min more specifically on unconstrained skills (d=.27, p < .0001). There was also some evidence that children who qualified for free-or-reduced-price lunch spent more time on instruction (5 min, d=.11, p < .05), specifically on unconstrained language

Table 2
Number of minutes observed in ISI main constructs.

	Prekindergarten				Kindergarten			
	(n=263 children)				(n=390 children)			
	Mean	SD	Min	Max	Mean	SD	Min	Max
Time on Instruction vs. *Other*								
Total observed time	185.07	62.98	23.95	312.55	219.25	59.18	43.43	306.00
Instruction	122.22	46.71	17.97	243.75	163.76	50.78	34.75	288.57
Management/routines	51.39	25.26	1.08	137.23	49.59	22.41	6.83	116.85
Off task	19.96	13.71	0.00	102.23	14.55	13.14	0.00	108.17
Primary Content of Instruction								
Language and Literacy Total	65.00	34.40	0.00	159.30	124.48	49.68	15.87	221.55
Constrained	*16.45*	*15.18*	*0.00*	*85.12*	*37.79*	*22.15*	*0.00*	*101.57*
Unconstrained	*34.57*	*19.65*	*0.37*	*101.80*	*41.29*	*25.20*	*0.00*	*139.55*
Math Total	35.56	26.55	0.00	145.57	43.44	26.04	0.00	116.40
Constrained	*21.21*	*15.79*	*0.00*	*78.50*	*20.04*	*12.97*	*0.00*	*57.48*
Unconstrained	*1.86*	*4.32*	*0.00*	*21.18*	*3.70*	*8.72*	*0.00*	*52.20*
Close-ended	*12.36*	*10.91*	*0.00*	*56.43*	*8.31*	*8.49*	*0.00*	*44.18*
Open and flexible-ended	*10.57*	*10.58*	*0.00*	*42.13*	*15.44*	*11.25*	*0.00*	*55.33*
Morning Meeting	32.59	24.15	0.00	114.80	25.70	20.85	0.00	99.52
Arts and Music	16.40	20.27	0.00	102.47	8.89	18.56	0.00	104.03
Science	3.77	9.76	0.00	54.28	4.88	13.76	0.00	81.18
Social Studies	1.84	8.38	0.00	63.65	0.39	3.83	0.00	40.93
Socio-Emotional Learning	0.44	2.14	0.00	28.93	0.20	1.30	0.00	10.23
Self-Regulation	0.53	1.99	0.00	16.28	0.34	2.21	0.00	38.90
Motor Development	1.14	3.44	0.00	40.20	1.16	4.92	0.00	32.75
Other/Unknown	28.00	32.88	0.00	173.25	11.66	18.62	0.00	115.08
Format								
Whole group	80.99	38.62	0.00	189.43	103.68	38.43	0.72	181.93
Small group	11.71	15.38	0.00	67.95	22.04	21.87	0.00	95.97
Centers	62.48	29.35	0.00	139.88	43.39	34.87	0.00	133.82
Individual	30.43	23.22	0.00	99.07	52.69	32.22	0.00	182.37

Note: Constrained and unconstrained skills do not equal to total time spent on language and literacy or math because we also coded time on planning and directions for these content areas as well (see Appendix A Tables 1 and 2). As described in the measures section, close-ended and open and flexible-ended are alternative math content measures for constrained/unconstrained. Math measures shown in italics do not sum to time spent on total math accordingly.

Table 3
Child-, teacher-, and school-level variation in learning opportunities.

	Prekindergarten (ICCs)			Kindergarten (ICCs)		
	Child-level	Teacher-level	School-level	Child-level	Teacher-level	School-level
Time on Instruction vs. *Other*						
Instruction	0.426	0.430	0.144	0.306	0.615	0.079
Management/routines	0.571	0.392	0.037	0.358	0.613	0.029
Off task	0.425	0.351	0.224	0.105	0.661	0.234
Primary Content of Instruction						
Language/literacy instruction						
Total	0.428	0.463	0.109	0.373	0.607	0.020
Constrained	0.292	0.672	0.036	0.303	0.697	0.000
Unconstrained	0.736	0.264	0.000	0.484	0.516	0.000
Math instruction						
Total	0.294	0.706	–	0.282	0.629	0.089
Constrained	0.199	0.544	0.257	0.212	0.683	0.105
Unconstrained	0.627	0.373	0.000	0.292	0.705	0.003
Close-ended	0.135	0.544	0.321	0.303	0.666	0.031
Open-ended	0.442	0.558	0.000	0.388	0.514	0.098
Format						
Whole group	0.084	0.916	0.000	0.032	0.889	0.079
Small group	0.224	0.776	0.000	0.242	0.756	0.002
Centers	0.101	0.793	0.106	0.120	0.880	–
Individual	0.088	0.727	0.185	0.190	0.810	0.000

skills (4 min, d=.22, $p < .05$). Black students spent more time in total language and literacy instruction (11 min, d=.38, $p < .01$), particularly on unconstrained literacy (7 min, d=.36, $p < .05$). Latino students too spent more time on literacy (7 min, d=.23, $p < .05$) than White children and again, specifically on unconstrained skills (7 min, d=.37, $p < .01$). They also spent less time on open-ended math instruction than White children (3 min, d= -.27, $p < .05$). Heterogeneity by child baseline skill level did not show a consistent pattern, and there were no statistically significant differences in learning experiences by dual language status.

Table 4
Variation in learning opportunities by prekindergarten student characteristics.

	DLL	Female	FRL	Black	Asian	Other	Latino	Low vocab	High vocab	Low math	High math
Time on Instruction vs. *Other*											
Instruction	-0.647	8.589***	5.112*	4.456	4.646	-2.356	1.445	2.977	-1.298	-3.063	-1.972
Management/routines	0.346	-7.188***	-3.475	2.502	1.277	4.738	-1.285	-4.515	0.168	-2.342	0.523
Off task	-0.004	-1.353	0.314	-3.097	-2.453	-1.952	2.438	1.874	-1.055	4.828**	-1.322
Primary Content of Instruction											
Language/literacy instruction											
Total	-1.282	7.159***	3.137	10.777**	5.173	1.391	6.604*	1.950	0.809	-0.806	-1.315
Constrained	-1.679	1.718	-1.835	3.845	3.936	0.182	-0.997	-0.471	2.060	-1.553	0.335
Unconstrained	-0.195	5.399**	4.254*	7.051*	0.712	1.379	7.197**	0.712	-1.352	1.211	-1.700
Math instruction											
Total	1.507	-0.383	0.844	-2.906	-2.958	-1.542	-2.962	-2.165	-2.844	-0.855	2.093
Constrained	1.486	0.247	1.061	-2.006	-2.465	-1.233	-2.246	-1.576	-2.064	0.915	2.009
Unconstrained	0.128	-0.327	-0.046	0.218	0.142	0.439	0.141	-0.618	-0.720	-0.796	0.273
Close-ended	1.316	0.260	-0.014	-1.031	0.449	-0.496	0.630	0.613	-1.341	1.599	1.512
Open-ended	0.406	-0.000	0.709	-0.884	-2.608	-0.226	-2.882*	-3.149**	-1.350	-1.819	0.848
Format											
Whole group	0.931	1.185	0.837	-0.151	3.459	0.312	1.304	1.982	0.243	1.472	0.549
Small group	-0.077	-0.781	0.718	2.739	-0.144	-1.634	-1.408	-0.787	-0.574	-0.778	-0.404
Centers	-0.055	-0.417	-0.721	3.488	-5.086	-0.529	0.594	-3.836*	-2.690	-1.598	-0.525
Individual	-0.820	-0.100	-0.715	-6.381*	-0.022	0.857	-1.018	2.503	2.977	1.730	0.796

Note: DLL=Dual Language Learner; FRL=Free-reduced-lunch eligible; Low vocab=scored in bottom quartile on PPVT fall assessment; High vocab=scored in top quartile on PPVT fall assessment; Low math=scored in bottom quartile on W-J Applied Problems fall assessment; High math=scored in top quartile on W-J Applied Problems fall assessment. Reference category for race/ethnicity is White. Reference category for pretest is middle two quartiles. Models also include random intercepts for schools and classrooms.

Table 5
Variation in learning opportunities by kindergarten student characteristics.

	Prek Attender	DLL	Female	FRL	Black	Asian	Other	Latino	Low vocab	High vocab	Low math	High math
Time on Instruction vs. *Other*												
Instruction	-0.394	2.247	4.788***	-0.015	-2.516	0.905	-1.912	-1.153	-1.721	-2.756	-0.140	0.026
Management/routines	-0.934	-0.180	-1.240*	0.205	0.724	-0.859	-0.286	2.275**	0.300	-0.537	0.708	-0.976
Off task	-1.653	-1.304	-2.517***	2.018**	1.999	-0.805	-2.144	0.321	2.960**	-0.253	1.279	-0.622
Primary Content of Instruction												
Language/literacy instruction												
Total	-0.534	-2.347	5.636***	-5.377**	-6.953*	-5.649	-6.258	-7.237**	0.234	4.766*	-3.476	-0.114
Constrained	1.546	0.267	0.607	-0.064	-1.191	0.909	-1.306	-0.899	-0.093	0.077	-2.362*	0.458
Unconstrained	-0.652	-3.467*	4.687***	-5.439***	-5.849*	-7.035**	-3.970	-5.537**	-0.666	6.052***	-2.516	-0.917
Math instruction												
Total	0.377	1.281	-1.297	0.509	-0.783	0.159	-0.151	-0.064	0.048	-2.101*	0.593	0.160
Constrained	-0.122	1.406	0.007	0.282	-0.508	0.332	0.436	0.198	0.051	-1.804*	0.877	0.861
Unconstrained	0.514	-0.228	-0.950*	0.582	0.142	0.405	0.367	0.103	0.318	0.383	-0.369	-0.982
Close-ended	0.016	0.280	0.044	-0.066	1.258	0.170	1.748	1.747**	0.307	-0.583	0.510	-0.071
Open-ended	0.628	0.768	-0.825	0.526	-1.657	0.155	-1.130	-1.524	-0.066	-0.752	0.067	0.131
Format												
Whole group	0.740	1.497	-1.140	0.089	2.601	-0.100	0.072	1.834	0.893	-1.950	2.185*	-0.494
Small group	-0.351	6.265***	0.066	0.104	-1.357	0.973	4.334	-0.021	0.304	-1.883	0.534	-1.893
Centers	-1.933	0.983	2.821**	1.098	0.930	3.782*	-3.589	-1.970	0.882	1.576	-3.294**	-0.289
Individual	1.825	-2.042	-2.802**	-0.436	-0.672	-2.250	-0.171	1.827	-1.324	0.877	1.593	3.104*

Note: Prek=BPS prekindergarten attender; DLL=Dual Language Learner; FRL=Free-reduced-lunch eligible; Low vocab=scored in bottom quartile on PPVT fall assessment; High vocab=scored in top quartile on PPVT fall assessment; Low math=scored in bottom quartile on W-J Applied Problems fall assessment; High math=scored in top quartile on W-J Applied Problems fall assessment. Reference category for race/ethnicity is White. Reference category for pretest is middle two quartiles. Models also include random intercepts for schools and classrooms.

The kindergarten findings most consistent with the prekindergarten results were for gender (see Table 5). Girls spent more time on instruction (5 min, d=.09, p < .001), less time off task (3 min, d= -.19, p < .01), and less time on management/routines (1 min, d=-.06, p < .05) than boys. The additional time on instruction for girls was concentrated in language and literacy (6 min total, d=.15, p < .001, of which 5 min was on unconstrained language and literacy, d=.19, p < .001). Girls spent more time in center activities (3 min, d=.08, p < .01), while boys spent more time on individual activities (3 min, d=.09, p < .01). Learning experiences were heterogeneous by free-or-reduced-price lunch status and race/ethnicity but in the opposite direction from prekindergarten. Children eligible for free-or-reduced-price lunch spent less time on language and literacy (specifically, unconstrained skill activities, 5 min, d=-.14, p < .001) than their peers. They were more off task than their peers as well (2 min, d= .15, p < .01). Black, Asian, and Latino children spent less time on unconstrained language and literacy than White children (unconstrained skill coefficients ranged from 6–7 min, d= -.22 to -.28, p < .05). Latino children also spent about 2 min more on close-ended math activities than White children (d=.21, p < .01). For child baseline skills, children with high vocabulary spent more time on unconstrained literacy skills (6 min, d=.24, p < .001) and slightly less time on math instruction (about 2 min, d=-.14, p < .01). Children with

Table 6
Relations between students' learning opportunities and gains in their prekindergarten and kindergarten skills.

	Prekindergarten (n=263 children)				Kindergarten (n=390 children)						
	PPVT	W-J AP	REMA	Digit Span	PPVT	W-J AP	REMA	Digit Span	DIBELS LNF	DIBELS PSF	DIBELS NWF-WWR
Time on Instruction vs. *Other*											
Instruction	0.031	0.007	0.035	0.005	0.001	0.006	0.001	-0.000	-0.038	-0.006	-0.036
Management/routines	-0.018	-0.000	-0.039	-0.004	-0.027	-0.013	0.017	-0.002	-0.020	-0.032	0.011
Off task	-0.182	-0.043*	-0.059	-0.011	-0.068	-0.009	-0.071**	-0.000	0.038	-0.006	0.023
Primary Content of Instruction											
Language/literacy instruction											
Total	0.041	-0.006	0.022	0.006*	-0.047	0.003	0.007	-0.002	0.019	0.010	-0.044*
Constrained	0.069	-0.014	-0.029	-0.002	-0.047	0.011	0.022	0.001	0.047	0.071	0.043
Unconstrained	-0.092	0.024	0.032	0.000	0.074	-0.008	-0.015	-0.002	-0.034	-0.083	-0.029
Math instruction											
Total	0.065	0.001	0.010	-0.001	-0.049	0.006	-0.010	-0.001	-0.019	0.089	0.069
Constrained	-0.150	-0.002	-0.007	0.005	-0.077	0.011	-0.012	0.001	-0.034	-0.119	0.051
Unconstrained	0.150	0.002	0.007	-0.005	0.086	-0.013	0.011	-0.002	-0.041	0.105	-0.027
Close-ended	-0.144	-0.033	-0.016	-0.008	-0.229*	0.013	0.006	0.002	0.092	-0.136	0.178*
Open-ended	0.180	0.016	0.020	0.004	-0.030	-0.004	-0.029	-0.003	-0.080	0.155	0.033
Format											
Whole group	0.015	0.009	0.015	0.006*	-0.078**	-0.010	-0.017	-0.000	0.025	-0.047	-0.029
Small group	0.081	-0.013	-0.023	0.001	-0.046	-0.005	-0.007	-0.000	-0.035	0.022	-0.034
Centers	-0.009	-0.010	0.017	-0.002	0.084***	0.012**	0.014	0.000	-0.039	-0.004	0.009
Individual	-0.047	0.008	-0.043*	-0.005	-0.014	-0.006	0.003	-0.000	0.023	0.015	0.016

Note: Models control for child race/ethnicity, gender, free/reduced lunch, dual language, child age, total observed time, interval between baseline and outcome measures, CLASS Instructional Support, number of observations per child, parent covariates, and the requisite fall assessment (for REMA, for the fall assessment, we controlled for W-J Applied Problems and for DIBELS, we controlled for both available fall tests). Models also include random intercepts for schools and classrooms. FSF = First Sound Fluency, LNF = Letter Naming, PSF = Phoneme Segmentation Fluency, NWF WWR = Nonsense Word Fluency Whole Word Reading.

low math scores spent less time on constrained literacy (2 min, d=.11, p < .05), less time in centers (3 min, d=-.09, p < .01), and more time in whole group (2 min, d=.06, p < p < .01), while children with high math scores spent more time in individual activities (3 min, d=.10, p < .05). Unlike in prekindergarten, there was some evidence of variation in learning experiences for DLLs (3 min less on language/literacy unconstrained skills, d=-.14, p < .05 and 6 min more in small groups, d= .29, p < .001) versus their monolingual peers. There was no evidence of variation in learning experiences by whether the child had attended Boston prekindergarten in the previous year.

9. Relations between learning experiences and child skill gains (RQ3)

In Table 6, we display unstandardized associations between our key ISI predictors and gains in children's outcomes, controlling for CLASS Instructional Support, children's baseline scores, and child and parent background characteristics. Overall, we found little evidence of consistent relations in either grade.

In prekindergarten, all relations were null with four exceptions. There was a small negative relation between time spent off-task and math gains as measured by the W-J Applied Problems only (-.12 *SD*s, p < .05). We also found evidence of a small positive, statistically significant relation between total time spent in language and literacy instruction and EF gains (.18 *SDs*; p < .05). For learning formats, we found a positive, statistically significant relation between time in whole group and EF gains (.22 *SDs*; p < .05) and a negative, statistically significant relations between time in individual instruction and math gains as measured by the REMA (-.11 *SDs*; p < .05). These findings may be spurious given the number of models fit (4 outcomes and 15 predictors, for a total of 60 prekindergarten models).

In kindergarten, we found mostly null results and only one finding was aligned with our prekindergarten findings. Specifically, time spent off-task was negatively associated with math gains, though for REMA and not W-J Applied Problems as in prekindergarten (-.12 *SDs*; p < .01). We also found evidence of a small negative, statistically significant relation between total time spent in language and literacy instruction and gains in Nonsense Word Fluency Whole Word Reading (-.16 *SDs*; p < .05). For learning formats, there was a negative, statistically significant relation between time in whole group and receptive vocabulary gains (-.12 *SDs*; p < .01) and a positive relation between time spent in centers and gains in receptive vocabulary (-.11 *SDs*; p < .001) and math gains as measured by the W-J Applied Problems (-.10 *SDs*; p < .01). There were two other statistically significant relationships, but they were not aligned by domain (i.e., more time close-ended math instruction related to receptive vocabulary). While cross-domain effects are possible and have some support in the literature (Weiland & Yoshikawa, 2013), these findings may be spurious given the number of kindergarten models fit (7 outcomes and 15 predictors, for a total of 105).

10. Robustness checks

We conducted sensitivity analysis for RQs 2 and 3. These checks are described in full in Appendix B. In brief, we found that RQ2 findings were largely robust to using either random intercepts for classrooms and schools or fixed intercepts for classrooms. For RQ3, we conducted eight robustness checks: (1) fixed effects for classrooms; (2) aggregating ISI constructs to the classroom level; (3) replacing CLASS Instructional Support control variable with the total CLASS score; (4) excluding CLASS scores; (5) using alternative scores for our outcome measures, where available; (6) imputing missing data for the DIBELS which was missing at higher rates than other study data; (7) operationalizing the ISI as a percentage of time observed; and (8) analytical choices in creating our language and literacy constructs that differed from some other ISI studies. For checks 1 and 2, only one-third of our original results were robust. For checks 3–6, the majority of our results held in terms of statistical significance and magnitude with a few exceptions that had similar magnitudes but were no longer statistically significant. For check 7, the majority of our results were robust in terms of statistically significance except for two that were no longer statistically significant and had decreased magnitude. For check 8, we replicated prior work with the ISI and found largely null results. Overall, our RQ3 robustness findings underscore results from our primary strategy – e.g., there is little evidence of consistent relations between ISI constructs and children's gains in prekindergarten and kindergarten.

11. Discussion

Using an extremely detailed, child-level prekindergarten and kindergarten classroom observational measure, we contribute new evidence to the next generation of measurement work in early education (Burchinal, 2018; Weiland, 2018). Using the ISI, we find some evidence supporting the hypothesis that classroom-level measures mask important variation in learning experiences across children enrolled in the same classroom, including for content and learning formats. We add new evidence to this literature (Connor et al., 2006; Sabol et al., 2018; Vitiello et al., 2012) showing the value of distinguishing *constrained* from *unconstrained* language, literacy, and math learning experiences specifically (Paris, 2005; Snow & Matthews, 2016). Our findings also highlight how child-level measures can capture differences in children's learning experiences by subgroups, something that classroom-level measures cannot do but that may be important in promoting greater equity in early learning opportunities and for improving outcomes for all children. However, despite the substantial variation in learning experiences across children enrolled in the same classroom, there are few associations with gains in children's language, literacy, math, and executive function skills in prekindergarten and kindergarten. These latter findings underscore the difficulty the field faces in developing measures that consistently predict children's learning gains.

12. Variation across individual children in the same classroom

Taking our major findings in turn, consistent with the masking hypothesis, we found considerable variation in learning experiences across children enrolled in the same classroom using the ISI. Again, this is variation that classroom-level measures cannot detect. Notably, there was more variation in both grades at the child level in language and literacy than in math. Interestingly, in prekindergarten but not kindergarten, results for *constrained* versus *unconstrained skill types* differed markedly. In prekindergarten, the majority of the variance in both unconstrained language and literacy and unconstrained math was at the child level. For constrained skills in prekindergarten, the majority of the variance was at the teacher level. In terms of format, teachers drove variation in both prekindergarten and kindergarten. These results underscore prior findings that learning experiences vary across young children in the same classroom (Connor et al., 2006; Vitiello et al., 2012), and add nuance to this evidence. That is, variation in children's learning experiences can differ by grade, content area, and skill type within content domains. Our findings reveal greater child autonomy in prekindergarten classrooms, with particular leeway to choose unconstrained language activities.

13. Subgroup findings

Ours is the first study to use the ISI to examine how learning experiences vary by student subgroup and we find that it is sensitive to detecting such differences. This too is a unique feature of child-level versus classroom-level measures. We detected some heterogeneity in learning experiences across student subgroups in prekindergarten, specifically on unconstrained literacy skills favoring girls, children from low-income families, Black students, and Latino students. Kindergarten findings were similar to prekindergarten findings for girls only and in the opposite direction for other groups for unconstrained literacy. In prekindergarten, we found little evidence of variation by child baseline skills and none by dual language learner status. In contrast, in kindergarten, high-vocabulary children spent more time on unconstrained literacy skills and slightly less time on math instruction, while dual language learners spent less time on unconstrained skills. Our findings for girls align with prior evidence using the Snapshot (Early et al., 2010) and our prekindergarten dual language learner findings align with evidence from the LISn (Bratsch-Hines et al., 2019).

The flip in direction of our findings between prekindergarten and kindergarten for Black students, Latino students, and students from low-income families merits further study. Differences are considerable across a month or school year (e.g., 11 min more on language and literacy instruction for Black versus White students in prekindergarten represents about 110 min per month). But these findings might partially explain why students from some subgroups appear to gain more from prekindergarten programs, yet may not maintain their gains in the elementary years (Phillips et al., 2017). These findings also underscore calls for more investigations generally into kindergarten teaching and learning contexts (Weiland, et al.,2021, In press). While prior literature highlights inequities in access to high-quality early learning experiences across programs (Chaudry et al., 2017; Latham et al., 2020), our findings point to the importance of investigating inequities *within* early education classrooms as well and the need for better understanding of the professional development necessary to disrupt inequities.

14. Predictive validity findings

Our most consistent predictive validity findings were for off-task behaviors, which showed small, negative relations with math gains in prekindergarten and math gains in kindergarten, though with different math measures in each grade. Notably, one prior ISI study found small, negative associations between children's off-task behaviors and their EF gains in kindergarten and in their reading comprehension gains in first grade (Moffett & Morrison, 2020). In addition, Sabol & colleagues (2018) found negative associations between a child-level measure of children's negative engagement in preschool and gains in their vocabulary, literacy, self-regulation, and teacher-child relationships. Interestingly, most of the variation in off-task behavior in our study was not at the child level, in either grade. And at the classroom level, the percent of time children were off task was most consistently and strongly correlated with the three CLASS domains. Together, child-level findings across these studies may point to a *malleable factor* for improving outcomes. That is, classroom and school contexts appear to exert considerable influence on children's off-task behaviors and perhaps by extension, their learning gains. For example, relatively simple changes to teacher classroom practices such as more predictable classroom routines and reduced time sitting still and listening may increase children's engagement and thus improve their learning in these important early years. Our off-task behavior findings did not hold across all robustness checks, however. Our findings need further replication to identify any possible implications.

But the null associations between most of our ISI measures and children's gains give pause in particular regarding the ISI's predictive validity and the masking hypothesis explanation for the weak predictive validity of classroom-level measures. There are several potential explanations for these findings. First, although the ISI captures time on content, it does not capture *depth* of content, nor *quality* of instruction. Presently, there are no systematic measures of the former in early childhood settings, though some experts have highlighted the importance of depth in supporting young children's learning (Neuman, 2014). There are measures assessing the quality of specific content delivery, though here too findings are mixed (Zaslow et al., 2016) and more measurement work is needed. Second, in prekindergarten, it is possible that children's experiences are too variable from day to day and that the ISI accordingly suffers from considerable measurement error. Many early education classrooms, including in the Boston Public Schools, prioritize student choice across activities, settings, and content areas. Even though we observed on two different days, it is possible that we would need to observe on additional days across a longer time period to obtain a representative sample of how individual prekindergarteners spend their time and to better approximate the cumulative effects of specific learning experiences. However, in kindergarten, less time was spent in centers – the key choice time in both grades – and we found no more evidence of predictive validity than in prekindergarten. In addition, given the resources required to code the ISI for two days, more observations are unlikely to be feasible for most teams. A third possibility is that our range was restricted. Time on instruction and content areas in our study compares favorably to other available benchmarks (e.g., Early et al., 2010) and most study classrooms used the Boston *Focus on Early Learning* curricula (McCormick et al., 2020). It is possible that in lower-quality settings with less consistent curricula there would be additional variability in children's experiences that would be predictive of children's gains. Finally, our study did not measure the degree to which instruction was differentiated appropriately for individual children's learning needs. This is another direction suggested by experts (Burchinal et al., 2018; Connor et al., 2020) that may hold promise.

From a practical perspective, the ISI is also very resource intensive. It requires videotaping, which many practitioners dislike. We found a 2.5 h video required about 8 h of coding time. The ISI's creators recognized this issue and developed one such new live-coding system which is currently undergoing testing and validation (Connor et al., 2020). This system might show better predictive validity as well, since it provides teachers with information to inform their differentiation of instruction for individual children. Increasingly, there are more ways in which to use cameras and microphones in classrooms to gather footage and new methods like machine learning that can be used to identify important patterns in large datasets (Weiland & Guerrero Rosada, 2022). This could include actionable, timely data to teachers about subgroup inequities like those we identified in the present study. This work is in its infancy but eventually could be highly scalable and more cost-effective than current methods.

15. Limitations

Our study has important limitations. First, it is observational, not causal. Second, external validity is limited to BPS classrooms which employ highly qualified teachers and in which most classrooms use the aligned Boston *Focus on Early Learning* curricula. Third, there are some demographic differences in our sample versus the broader population of BPS prekindergarten and kindergarten students. For example, on average, 65% of district prekindergarteners and 68% of district kindergarteners were eligible for free-reduced-priced lunch, versus 58% and 61% of our sample, respectively. Study results may not generalize to all district students in these grades. Fourth, to maximize statistical power, we included all available children in our prekindergarten and kindergarten samples. Following the same children from prekindergarten to the end of kindergarten might have led to different findings, a topic we plan to take up in a future study. Fifth, for subgroup analysis, not all classroom-level sample members included all subgroups and their relevant counterparts (e.g., a classroom might have had Black and White children but not other racial/ethnic groups). Accordingly, not all classrooms contributed to all subgroup estimates. This is a difficult problem to overcome given enrollment patterns but nonetheless important in interpreting our subgroup findings. Sixth, some recent studies have found that child-level behavior measures analyzed at the child or classroom predict preschoolers' gains (Burchinal et al., 2021; Justice et al., 2018; Pianta et al., 2020). Magnitudes are modest and findings are inconsistent but it is possible that our findings may have differed with a count approach. This is a direction for future research with the ISI tool. Relatedly, we found somewhat different results when aggregating the ISI to the classroom level (Appendix B Table 7). This is a topic for future research; it is possible that there is less measurement error, for example, at the classroom level. Seventh, studies that use other child-level measures may find more consistent predictive validity with child gains. That is, ours is a limited test of the masking hypothesis with one measure rather than a definitive test. Eighth, our study did not include socio-emotional measures, captured only one element of language development (receptive vocabulary), and did not include literacy measures in prekindergarten. A broader range of child skills would have enhanced our study's contribution. Finally, it is possible our findings accordingly are limited by our statistical power. We do note however that our sample is aligned with that of most ISI studies and also that the measure is very intensive to code. A larger sample is difficult in terms of resources and coding, though would be instructive for the field.

16. Conclusion

Despite these limitations, our study contributes new evidence to the next generation of measurement work in early childhood education (Burchinal, 2018; Weiland, 2018). Consistent with emerging evidence, we find that learning experiences vary substantially across individual young children enrolled in the same classroom and across student subgroups – variation that is masked by classroom-level observational measures and that may be key to promoting equity in children's learning experiences and outcomes in the early grades. However, while children's off-task behaviors showed some evidence of predictive validity (for math only), most ISI constructs did not. Thus, the masking hypothesis appears to have some empirical evidence but measurement at the child-level at least with the ISI does not appear to be a silver bullet to solving the widely noted predictive validity issues of classroom-level classroom measures. Given the importance of observational measures to policy-level efforts focused on improving children's early learning gains (Bassok et al., 2019; U.S. Department of Health and Human Services, 2018) and the large, consequential gaps in school readiness between more advantaged and less advantaged groups (Chaudry et al., 2017), additional measurement work is needed in other contexts.

Data availability

The authors do not have permission to share data.

CRediT authorship contribution statement

Christina Weiland: Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Lillie Moffett:** Conceptualization, Software, Investigation, Visualization, Supervision, Writing – original draft, Writing – review & editing. **Paola Guerrero Rosada:** Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft, Writing – review & editing. **Amanda Weissman:** Software, Validation, Formal analysis, Investigation, Data curation, Vi-

sualization, Writing – original draft, Writing – review & editing. **Kehui Zhang:** Software, Formal analysis, Investigation, Writing – review & editing. **Michelle Maier:** Conceptualization, Funding acquisition, Writing – review & editing. **Catherine Snow:** Conceptualization, Project administration, Funding acquisition, Writing – review & editing. **Meghan McCormick:** Conceptualization, Investigation, Supervision, Project administration, Funding acquisition, Writing – review & editing. **JoAnn Hsueh:** Conceptualization, Supervision, Project administration, Funding acquisition, Writing – review & editing. **Jason Sachs:** Conceptualization, Project administration, Funding acquisition, Writing – review & editing.

Funding

This research was conducted as a part of a study funded by 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. Lillie Moffett's and Amanda Weissman's work was supported by the Institute of Education Sciences Predoctoral Fellowship at the University of Michigan (Grant R305B150012). And Lillie Moffett is further supported by the NSF Graduate Research Fellowship (Grant 1650114).

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, Deborah Ball, the MDRC team (Jennifer Yeaton, Mirjana Pralica, Rama Hagos, 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), Frederick Morrison, and the late Carol Connor for advising our team on the ISI Measure.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ecresq.2022.11.008.

References

Al Otaiba, S., Connor, C., Lane, H., Kosanovich, M. L., Schatschneider, C., & Wright, T. L (2008). Reading First kindergarten classroom instruction and students' growth in phonological awareness and letter naming-decoding fluency. *Journal of School Psychology, 46*, 281–314.

Alvidrez, J., & Weinstein, R. S. (1999). Early teacher perceptions and later student academic achievement. *Journal of Educational Psychology, 91*(4), 731–746.

Ansari, A., & Purtell, K. M. (2017). Activity settings in full-day kindergarten classrooms and children's early learning. *Early Childhood Research Quarterly, 38*, 23–32.

Bassok, D., Dee, T. S., & Latham, S. (2019). The effects of accountability incentives in early childhood education. *Journal of Policy Analysis and Management, 38*(4), 838–866.

Bratsch-Hines, M. E., Burchinal, M., Peisner-Feinberg, E., & Franco, X. (2019). Frequency of instructional practices in rural prekindergarten classrooms and associations with child language and literacy skills. *Early Childhood Research Quarterly, 47*, 74–88.

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

Burchinal, M., Garber, K., Foster, T., Bratsch-Hines, M., Franco, X., & Peisner-Feinberg (2021). Relating early care and education quality to preschool outcomes: The same or different models for different outcomes? *Early Childhood Research Quarterly, 55*, 35–51.

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

Chien, N. C., Howes, C., Burchinal, M., Pianta, R. C., Ritchie, S., Bryant, D. M., & Barbarin, O. A. (2010). Children's classroom engagement and school readiness gains in prekindergarten. *Child Development, 81*, 1534–1549.

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

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 math assessment. *Educational Psychology, 28*, 457–482.

Connor, C. M., Adams, A., Zargar, E., Wood, T. S., Hernandez, C., & Vandell, D. L. (2020). Observing individual children in early childhood classrooms using Optimizing Learning Opportunities for Students (OLOS): A feasibility study. *Early Childhood Research Quarterly, 52B*, 74–89.

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

Connor, C. M., Morrison, F. J., & Slominski, L. (2006). Preschool instruction and children's emergent literacy growth. *Journal of Educational Psychology, 98*, 665–689.

Connor, C. M., Ponitz, C. C., Phillips, B. M., Travis, Q. M., Glasney, S., & Morrison, F. J. (2010). First graders' literacy and self-regulation gains: The effect of individualizing student instruction. *Journal of School Psychology, 48*, 433–455.

Connor, C. M., Spencer, M., Day, S. L., Giuliani, S. I., McLean, L, & Morrison, F. J. (2014). Capturing the complexity: Content, type, and amount of instruction and quality of the classroom learning environment synergistically predict third graders' vocabulary and reading comprehension outcomes. *Journal of Educational Psychology, 106*, 762–778.

Connor, C. M. D., Morrison, F. J., & Katch, L. E. (2004). Beyond the reading wards: Exploring the effect of child-instruction interactions on growth in early reading. *Scientific Studies of Reading, 8*, 305–336.

Day, S. L., Connor, C. M., & McClelland, M. M. (2015). Children's behavioral regulation and literacy: The impact of the first grade classroom environment. *Journal of School Psychology, 53*, 409–428.

Duncan, S. E., & DeAvila, E. A. (1998). *PreLAS*.

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

Early, D. M., Iruka, I. U., Ritchie, S., Barbain, O., Winn, D. M. C., Crawford, G. M., & Bryant, D. M. (2010). How do pre-kindergarteners spend their time? Gender, ethnicity, and income as predictors of experiences in pre-kindergarten classrooms. *Early Childhood Research Quarterly, 25*, 177–193.

Fuligni, A. S., Howes, C., Huang, Y., Hong, S. S., & Lara-Cinisomo, S. (2012). Activity settings and daily routines in preschool classrooms: Diverse experiences in early learning settings for low-income children. *Early childhood research quarterly, 27*(2), 198–209.

Gehlbach, H., & Robinson, C. D. (2018). Mitigating illusory results through preregistration in education. *Journal of Research on Educational Effectiveness, 11*(2), 296–315.

Gilliam, W. S., Maupin, A. N., & Reyes, C. R. (2016). Early childhood mental health consultation: results of a statewide random-controlled evaluation. *Journal of the American Academy of Child and Adolescent Psychiatry, 55*(9), 754–761.

Good, R. H., Kaminski, R. A., Cummings, K. D., Dufour-Martel, C., Petersen, K., Powell-Smith, K., & Allin, J. (2011). *DIBELS next assessment manual*. Dynamic Measurement Group.

Gormley, W. T., Jr,, Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal pre-K on cognitive development. *Developmental Psychology, 41*(6), 872.

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

Guo, Y., Justice, L. M., Kaderavek, J. N., & McGinty, A. (2012). The literacy environment of preschool classrooms: Contributions to children's emergent literacy growth. *Journal of Research in Reading, 35*(3), 308–327.

Harms, T., Clifford, R. M., & Cryer, D. (1998). *Early childhood environment rating scale* (p. 10027). Teachers College Press Columbia University.

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

Justice, L. M., Jiang, H., & Strasser, K. (2018). Linguistic environment of preschool classrooms: What dimensions support children's language growth? *Early Childhood Research Quarterly, 42*, 79–92.

Latham, S., Corcoran, S.P., Sattin-Bajaj, C., & Jennings, J.L. (2020). Racial disparities in Pre-K quality: Evidence from New York City's universal Pre-K program. Working Paper 20-248. Providence, RI: Brown University, Annenberg Institute. https://www.edworkingpapers.com/sites/default/files/ai20-248.pdf.

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

Marks, R., Ellis, A., Moffett, L., Stilwell, S., & Morrison, F. J. (2016). *Individualizing student instruction classroom observations coding manual for schooling effects of executive function on academic achievement*. Frederick J. Morrison Lab Amended from Connor, C.M., Piasta, S., Al Otaiba, S., Day, S., Morrison, F.J., & Cameron, C. Version 80.06.03.2014.

McCormick, M., Hsueh, J., Weiland, C., & Bangser, M. (2017). *The challenge of sustaining preschool impacts: Introducing ExCEL P-3, a Study from the Expanding Children's Early Learning Network*. New York, NY: MDRC https://files.eric.ed.gov/fulltext/ED575653.pdf.

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.

McCormick, M., Weiland, C., Hsueh, J., Pralica, M., Weissman, A., Moffett, L., Snow, C., & Sachs, J. (2021). Is skill type the key to the PreK fadeout puzzle? Differential associations between enrollment in PreK and constrained and unconstrained skills across kindergarten. *Child Development*.

McCormick, M. P., Pralica, M., Weiland, C., Hsueh, J., Moffett, L., Guerrero-Rosada, P., & Sachs, J. (2022). Does kindergarten instruction matter for sustaining the prekindergarten (PreK) boost? Evidence from individual-and classroom-level survey and observational data. *Developmental Psychology*.

Moffett, L., & Morrison, F. J. (2020). Off-task behavior in kindergarten: Relations to executive function and academic achievement. *Journal of Educational Psychology, 112*, 938–955.

Muller, U., Kerns, K. A., & Konkin, K. (2012). Test–retest reliability and practice effects of executive function tasks in preschool children. *The Clinical Neuropsychologist, 26*, 271–287.

Neuman, S. B. (2014). Content-Rich Instruction in Preschool. *Educational Leadership, 72*, 36–40.

Noldus Information Technology. (2013). *Noldus observer xt*. Noldus Information Technology [Software] Available at http://www.noldus.com/human-behavior-research/products/the-observer-xt .

Nores, M., Friedman-Krauss, A., & Figueras-Daniel, A. (2022). Activity settings, content, and pedagogical strategies in preschool classrooms: Do these influence the interactions we observe? *Early Childhood Research Quarterly, 58*, 264–277.

Paris, S. G. (2005). Reinterpreting the development of reading skills. *Reading Research Quarterly, 40*, 184–202.

Phillips, D. A., Lipsey, M., Dodge, K., Haskins, R., Bassok, D., Burchinal, M., & Weiland, C. (2017). *Puzzling it out: the current state of scientific knowledge on pre-kindergarten effects*. Washington, DC: Brookings.

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

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring systemtm: manual K-3*. Baltimore: Paul H Brookes.

Pianta, R. C., Whittaker, J. E., Vitiello, V., Ruzek, E., Ansari, A., Hofkens, T., & DeCoster, J. (2020). Children's school readiness skills across the pre-k year: Associations with teacher-student interactions, teacher practices, and exposure to academic content. *Journal of Applied Developmental Psychology, 66*, 1–10.

Puma, M., Bell, S. H., Cook, R., & Heid, C. (2010). *Head start impact study: final report*. Washington, DC: U.S. Department of HHS, Administration for Children and Families.

Robinson-Cimpian, J. P., Lubienski, S. T., Ganley, C. M., & Copur-Gencturk, Y. (2014). Teachers' perceptions of students' mathematics proficiency may exacerbate early gender gaps in achievement. *Developmental Psychology, 50*, 1262.

Sabol, T. J., Bohlmann, N. L., & Downer, J. T. (2018). Low-income ethnically diverse children's engagement as a predictor of school readiness above preschool classroom quality. *Child Development, 89*, 556–576.

Sameroff, A. (2009). *The transactional model of development: How children and contexts shape each other* (pp. 3–21). American Psychological Association.

Sarama, J., & Clements, D.H. (2009). Manual for classroom observation (COEMET)-Version 3. Unpublished version.

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

Schochet, P. Z. (2009). An approach for addressing the multiple testing problem in social policy impact evaluations. *Evaluation Review, 33*(6), 539–567.

Schrank, F. A., McGrew, K. S., Ruef, M. L., & Alvarado, C. G. (2005). *Overview and technical supplement (Batería iii Woodcock- Muñoz Assessment service bulletin no.1)*. Riverside.

Smith, M. W., Brady, J. P., & Anastasopoulos, L. (2008). User's guide to the early language & literacy classroom observation Pre-K tool (ELLCO Pre-K). *Education Review*.

Snow, C. E., & Matthews, T. J. (2016). Reading and language in the early grades. *Future of Children, 26*, 57–74.

U.S. Department of Health and Human Services. (2018). *Use of classroom assessment scoring system (CLASS®) in head start*. U.S. Department of Health and Human Services Retrieved February 32019 from https://eclkc.ohs.acf.hhs.gov/designation-renewal-system/article/useclassroom-assessment-scoring-system-classr-head-start .

Vitiello, V. E., Booren, L. M., Downer, J. T., & Williford, A. P. (2012). Variation in children's classroom engagement throughout a day in preschool: Relations to classroom and child factors. *Early Childhood Research Quarterly, 27*, 210–220.

Vygotsky, L. (1978). *Mind and society: the development of higher mental processes*. Harvard University Press.

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

Weiland, C. (2018). Commentary: pivoting to the "how": Moving preschool policy, practice, and research forward. *Early Childhood Research Quarterly, 45*, 188–192.

Weiland, C., & Guerrero-Rosada, P. (2022). Widely-used measures of Pre-K classroom quality: what we know, Gaps in the field, and promising new directions. *Measures for early success: Supporting early learners and educators with innovative, equitable assessments*. MDRC. https://www.mdrc.org/sites/default/files/Widely_Used_Measures.pdf

Weiland, C., Ulvestad, K., Sachs, J., & Yoshikawa, H. (2013). Associations between classroom quality and children's vocabulary and executive function skills in an urban public prekindergarten program. *Early Childhood Research Quarterly, 28*, 199–209.

Weiland, C., Unterman, R., & Shapiro, A. (2021). The kindergarten hotspot: Literacy skill convergence between Boston Prekindergarten enrollees and non-enrollees. *Child Development*.

Weiland, C., Unterman, R., Shapiro, A., Staszak, S., Rochester, S., & Martin, E. (2020). The effects of enrolling in oversubscribed prekindergarten programs through third grade. *Child Development, 91*(5), 1401–1422.

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*, 311–333.

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

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

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

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