Identifying Actionable Classroom and Program Features for Scaling High-Quality Prekindergarten

by

Paola Andrea Guerrero-Rosada

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Education and Psychology)
in the University of Michigan
2023

Doctoral Committee:

Associate Professor Christina Weiland, Chair
Professor Nell Duke
Professor Ioulia Kovelman
Dr. Meghan McCormick, Manpower Demonstration Research Corporation
Professor Catherine Snow, Harvard Graduate School of Education

Paola Andrea Guerrero-Rosada

pguerre@umich.edu

ORCID iD: 0000-0002-0310-5890

Dedication

I dedicate this dissertation to each person who has taught me to challenge the odds. To my parents, Mauricio, and Samuel, especially.

Acknowledgements

My dissertation was conducted as a part of two studies. First, the Boston P-3: Identifying Malleable Factors for Promoting Student Success 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. Second, the “Building a Strong and Equitable Mixed-Auspice Prekindergarten System (MAPS): A Research-Practice Partnership to Support Boston’s Historic UPK Expansion” funded by the Heising-Simons Foundation. My work was supported by the National Academy of Education - Spencer Foundation, the University of Michigan Rackham Predoctoral Fellowship, and the Society for Child Development Student and Early Career Council Dissertation Research Funding Award. I appreciate the contributions of the Boston Public Schools Department of Early Education and Universal Prekindergarten Program (particularly TeeAra Dias, Jason Sachs, Yuzhu Xia), MDRC (particularly Mirjana Pralica and Michelle Maier), Harvard Graduate School of Education (Rachel Kushner and Louisa Penfold), and the University of Michigan team (particularly Anne Taylor, Amanda Weissman, Lillie Moffett, and Christina Daniel).

I want to thank each of the members of my committee for their important contributions to my training. Chris, I cannot thank you enough for your thoughtful mentorship, kindness throughout the process, and for helping me find an academic identity and voice. Catherine, thank you for your support and guidance, working with you has been one of the most formative and enjoyable experiences I have had. Meghan, thank you for sharing your impressive expertise in the field and providing valuable feedback in methods, writing, and dissemination. Nell, thank

you for helping me think deeply about early education practices, and for challenging my thinking through coursework, collaborations, and conversations. Ioulia, thank you for the opportunity to engage with the "En Nuestra Lengua" community and for modeling successful ways of navigating academia for international scholars.

I also want to thank the Equity in Early Learning (EEL) Lab (Amanda, Anna, Gloria, Tiffany, Jordy, Eleanor, Lillie, and Emily) and my cohort/friends at the University of Michigan (Mayra, Andrea, Jessica M, Sarah S, Jessica K, and Bernardette). Your good fellowship during these years—which included a pandemic and events of social unrest in Colombia and the U.S.—permanently motivated me and inspired me.

Thanks to my mentors and research team at Uniandes. Everything I learned through our projects, training, and policy interventions gave me strength when I ventured to learn about new contexts. Carolina, Andres, Eduardo, and members of the CM Lab, you will always be my first academic home.

Finally, I want to thank my family and friends for unconditionally supporting my decision to pursue a Ph.D. abroad—and trusting that I could do it successfully. Olga, Harold, Mauricio, Samuel, and others in my family: you have always given me courage to follow my dreams. Tat, Mar, Vivi, Roger, Ivan and Lorena: thank you for understanding what research means to me and keeping by my side during this process.

Table of Contents

List of Tables

List of Figures

List of Appendices

Abstract

In this dissertation, I provide actionable evidence on three related pressing questions for the early education field: what instructional features of classrooms predict children's academic gains in the prekindergarten year, whether the complexity of instruction from prekindergarten through first grade is aligned and contributes to children's within-year academic gains; and how to optimize the selection of centers participating in universal prekindergarten programs diminishing the risks of unintended patterns that could affect the quality of expansion programs. I use descriptive, psychometric, predictive, and geospatial methods to answer my research questions. In my first study, I focused on time use in prekindergarten classrooms. I compared the measurement properties of two instructional quality observational instruments: the Narrative Record (Farran et al., 2015) and the Individualizing Student Instruction (Connor et al., 2009) systems. Findings show that the NR and the ISI produce partially different descriptions of time use, and their resultant classroom instructional profiles are inconsistent. Although these profiles do not predict children's language and working memory gains, they do predict math gains with opposite directions depending on the used measure. I also discuss demographic differences by classroom type and illustrate a modeling approach that accounts for such differences in predictive models. My results can aid practitioners in monitoring equitable instruction across classrooms.

In my second study, I used multi-level linear regression models to identify the degree of instructional alignment in relation to children's exposure to content complexity and examine its contribution to children's within-grade gains in language (Dunn & Dunn, 2007) and math

(Clements et al., 2008; Woodcock et al., 2001, 2005). Findings show within-grade variation in the complexity of content instruction across classrooms. Moreover, I show that children in prekindergarten classrooms are exposed to highly complex content, and such complexity is not sustained but rather decreases in subsequent grade levels in reference to national and state learning standards. Parametric measures show consistent patterns, demonstrating that children are exposed to content of roughly the same complexity from prekindergarten to first grade. Although such variation does not predict children's within-grade language and math gains, these results highlight the importance of assessing instructional alignment and identifying its potential contribution to children's fade-out or convergence of skills after prekindergarten.

In my third study, I identified statistical and geo-spatial differences between centers that self-select to participate in the Boston Universal Prekindergarten (UPK) program and other centers in the Boston area, using administrative data from their licensing, quality rating and improvement system, and accreditation status. Results show that UPK appliers are located in and serve similar communities to non-appliers but are more likely to receive subsidies and participate in Quality Rating and Improvement Systems (QRIS). Differential participation in QRIS between appliers and non-appliers increases when models are restricted to CBOs receiving subsidies. These findings highlight the importance of monitoring quality at the population level using strategies independent of monetary incentives to secure equitable access to high-quality settings for low-income families.

Together, these three studies contribute to the discussion in the field about how to operate and scale high-quality public prekindergarten programs. First, by monitoring disparities in access to varied and complex instructional experiences across classrooms and schools. Second, by

implementing strategies that include population-level information to maximize equity and quality in selecting partners for universal prekindergarten programs.

Introduction

Achieving and sustaining quality as prekindergarten programs go to scale is a long-standing problem in early education, with many practical implications for policy makers and practitioners (Barnett et al., 2021). Significant public dollars are directed to this issue through helping programs adhere to state quality standards, participate in quality rating and improvement systems, and meet national quality benchmarks (Friedman-Krauss et al., 2019). Importantly, many school districts rely on research to make decisions about their programs that are likely to affect children's educational trajectories (Bardige et al., 2018). However, as a field, we are still learning what approaches, tools, and systems are optimal in achieving and sustaining strong positive effects on children's early learning (Weiland, 2018).

My dissertation consists of three related, yet standalone papers, aimed at examining actionable features of quality at the classroom, school, and system level in the context of a long-standing research-practice partnership. In study 1, I examine whether time use measures generate consistent descriptions of classroom instruction and consistent time use latent profiles. Further, I examine whether time use profiles predict children's gains in prekindergarten. In study 2, I examine whether the complexity of content instruction changes from prekindergarten to first grade and is associated with within-year language and math gains. To do so, I use a conceptually based measure representing classrooms' average grade level instruction and an empirically based parametric measure representing classrooms' content complexity. In study 3, I examine the differences between community-based organizations that apply to participate in the Boston

Universal Prekindergarten Program and those that do not, using statistical and geo-spatial methods.

My research questions for each study are:

Study 1: “Better Luck Next Time: Prekindergarten Classroom Time Use Profiles Are Measure-Dependent”

1. Do the NR and ISI produce similar time use descriptions of BPS Prekindergarten classrooms?
2. How do time use profiles based on the NR and the ISI compare in terms of the number of profiles derived by each instrument, their description, and potential differences in children and parent demographics, classroom quality, and classroom composition?
3. Do time use profiles of instructional content and formats based on the NR and the ISI predict children’s gains in language, math, and working memory, above and beyond differences in profile membership such as classroom quality and demographic composition?

Study 2: “Teacher-Reported Complexity of Instruction Does Not Predict Children’s Within-Grade Math and Language Gains in Prekindergarten, Kindergarten, or First Grade”

1. What is the average complexity of language/literacy and math instruction in prekindergarten, kindergarten, and first grade? How does it differ using conceptually based versus empirically based measures?
2. Do conceptually and empirically based measures of content complexity predict children’s within-grade prekindergarten, kindergarten, and first grade language and math gains?

Study 3: “Appliers to Mixed-Delivery Universal Prekindergarten Differ from Non-Appliers in Subsidy Receipt and QRIS Participation: Evidence from the Boston’s UPK Expansion”

1. Do community-based organizations applying to Boston UPK differ from non-appliers in terms of their capacity, structural quality, and the demographic characteristics of the communities where they are located?
2. Among centers receiving subsidies, do Boston UPK appliers differ from non-appliers in their capacity, structural quality, and the demographic characteristics of the children they serve?
3. Do proxies of structural quality from Boston community-based centers vary across census block groups and neighborhoods?

The contribution of my dissertation is twofold. Substantively, I found within-grade variation in classrooms’ instructional features, namely, how time is used, and the level of content complexity reported by teachers. These features of instructional quality are generally not assessed in mainstream quality measures. Although this variation fails to consistently predict children’s language, math, and working memory gains, I discuss alternative modeling approaches to be explored in larger samples. At a system level, I identified other differences indicative of centers’ financial and operational models, namely QRISs and subsidy participation. Methodologically, I identified differences in descriptive and predictive properties of time use measures, propose novel approaches to assess content complexity in early grades, and illustrate how using geos-spatial methods can help universal prekindergarten programs to identify underserved communities. However, I was unable to assess instructional differences between Boston UPK appliers and non-appliers due to a lack of variability in licensing indicators.

Chapter 1 - Study 1: Better Luck Next Time: Prekindergarten Classroom Time Use Profiles Are Measure-Dependent

Research has shown that prekindergarten programs tend to improve on the aspects of quality measured in large-scale systems (Bassok et al., 2017). However, although there is some consensus about how to assess the structural (administratively regulated aspects) and process (classroom interactions) features of prekindergarten quality (Burchinal, 2017; Mashburn et al., 2008; Pianta et al., 2016; Zaslow, 2011), existing measures of these components consistently show mixed or null associations with children's development (Brunsek et al., 2017; Guerrero-Rosada et al., 2021; Perlman et al., 2016; Weiland et al., 2013). This pattern of findings has spurred calls for additional measurement work in early education, emphasizing practical, scalable measures that localities can use to monitor and incentivize quality improvement (Weiland & Guerrero-Rosada, 2022).

Classroom time use is one area of quality measurement that several recent studies have begun to focus on, due to its potential to signal instructional differences across classrooms and predict prekindergartners' academic gains (Bratsch-Hines et al., 2019; Cabell et al., 2013; Chien et al., 2010; Early et al., 2010; Pianta et al., 2018; Weiland et al., 2023). However, available time use measures differ in several ways, with implications for descriptive and predictive research. First, measures vary in how learning activities and settings are defined, whether the measure focuses on experiences of the group or individual children, and whether data collection protocols involve using time-sampling strategies. These differences make it challenging to understand time use patterns across prekindergarten classrooms. Second, time use is intrinsically linked to important features of classroom quality, such as the curriculum being implemented and teachers' organizational strategies. This is a potential reason why, although time use measures adequately capture variation across classrooms (Bratsch-Hines et al., 2019; Burchinal et al., 2021; Justice et al., 2021; Nores et al., 2022; Weiland et al., 2023), associations with children's gains are not consistent across studies.

In the current study, we add to this literature by *exploring the descriptive and predictive properties of two measures of time use in public prekindergarten*: the Narrative Record (NR; Farran et al., 2015) and the Individualizing Student Instruction (ISI; Connor et al., 2009) systems, the first collected at the classroom level and the second collected at the child level but aggregated to the classroom level in this study. Our research team coded both measures from the same video recordings of Boston Public Schools (BPS) prekindergarten classrooms collected during the 2016-2017 school year. By coding and analyzing the same observational periods with two measures, we hold constant factors like length of observation, curricula, instructional content during the coding period, and sample characteristics. Accordingly, we can attribute any divergent results to differences in the measures we used.

We first compare descriptive statistics across the measures and test similarities across the resultant classroom time use profiles. Then, we examine whether instructional profiles across the measures vary by classroom quality and demographic makeup. This analysis helps to inform whether time use profiles are capturing the differences in instructional characteristics of interest instead of other classroom features such as its quality or demographic composition. Finally, we explore whether time use profiles consistently predict children's academic gains in prekindergarten and whether predictive power varies across measures. To our knowledge, this is the first study to directly compare two observational time use measures in early education settings. Our findings add to the broader literature on this highly actionable classroom feature and provide practical guidance on the trade-offs of using time measurement tools.

Why Measure Time Use in Prekindergarten Classrooms?

Although there is agreement that high quality early childhood education promotes positive child development, there is a lack of consensus on the active ingredients of classroom experiences that best promote children's outcomes. New directions in measurement have posited that the current dominant conceptual framework in the field that focuses on process quality – or teacher and child *interactions* – as the key active ingredient leaves out other factors driving children's gains, especially *content* of instruction (Maier et al., 2020). Measuring time use in early education settings has potential for capturing whether and how teachers expose children to high-quality instructional content using developmentally appropriate learning formats. However, it has been less frequently used despite addressing a critical gap in the literature, namely, classroom instructional features.

Work examining instructional aspects of process quality suggests that these dimensions are more predictive of children's academic and cognitive outcomes than relational aspects of process quality (Maier et al., 2020). Consistently, we propose that, if time use measures are reliable, examining classrooms' distinct instructional profiles in this way can shed light on *what* children are taught and whether there are important disparities in the amount of time children from different groups are exposed to foundational literacy, vocabulary, and math knowledge, rich scientific and social studies content, engagement with the arts, socio-emotional and motor development, among other prekindergarten experiences. Similarly, children's exposure to varied learning settings where they can exert autonomy, choice, strengthen attention, and interact with peers are essential to the development of higher-order skills (Lerkkanen et al., 2016).

Interestingly, researchers have also shown that time use is associated with teachers' process quality as measured by the Classroom Assessment Scoring System (CLASS PreK; Pianta et al., 2008) (Cabell et al., 2013; Nores et al. 2022). For example, in a sample of 314 classrooms that were part of a multi-site, randomized controlled trial (National Center for Research on Early Childhood Education —NCRECE; Phase II; Pianta et al., 2008), instructional support was highest during science activities (mean = 2.94, on a scale of 1 to 7), with lower scores during shared reading, social studies, literacy, and math (mean scores ranging from 2.33 to 2.62). Researchers observed the lowest levels of instruction support during art (Mean = 2.11) and cycles when no learning activities were taking place (Mean = 1.97). Another study conducted with a sample of 264 classrooms in New Jersey and Philadelphia found that there were higher levels of emotional support and classroom organization in classrooms that spent more time in literacy activities and less time in science and social studies. Classrooms with higher levels of classroom organization also spent more time on math instruction, on average (*b* coefficients between 0.14 and 0.15) (Nores et al., 2022). Nores and colleagues further found associations between time in learning settings and quality scores. Specifically, more time in choice activities and less time in whole group were associated with higher level of instructional support (*b* coefficients = 0.23 and -0.15 respectively).

Time use measures that can reliably describe the instruction that is happening in classrooms and further predict children's academic gains can be used to identify valuable and actionable classroom-level time use profiles programs can use to support quality improvement through training and coaching. Below, we present descriptive evidence from the literature on how time use has been measured across prekindergarten classrooms, and how such variation relates to children's developmental gains.

Describing Time Use Variation in Prekindergarten Classrooms

Describing time use in prekindergarten classrooms poses conceptual and methodological challenges. Prekindergarten classrooms are complex systems where many interactions—between students themselves and between students and teachers—take place simultaneously. Thus, generating consensus on what features should be described and further collecting, coding, and analyzing detailed data necessary to obtain reliable measures of such features is resource intensive. The field has approached the analysis of time use in prekindergarten classrooms in two ways: a variable-centered approach, where time spent on each instructional content area (e.g., math, language, literacy, science) and learning format (e.g., whole group, small group, individual learning) is measured independently from each other; and a person-centered approach, where groups of classrooms are described by their shared characteristics, namely, identifying instructional profiles (e.g., classrooms dedicating a large proportion of time to language and literacy instruction, classrooms dedicating a larger proportion of time to free-play, etc.).

Proportion of Time Spent in Content Areas and Learning Formats (Variable-Centered Approach)

When using a variable-centered approach, researchers have found systematic differences in the proportion of time that prekindergarten classrooms allocate to varied learning activities (e.g., language, literacy, and math instruction) and spend in different learning formats (i.e., whole-class, small groups, centers). In Table 1.1, we summarize findings from 10 previous studies of time use in prekindergarten classrooms applying this approach. Specifically, we show the proportion of time spent in learning settings and activities across studies, as well as details on the sample and instrument used in extant work. Findings from most studies in this area suggest that children spend about a third of their day in whole-group or large-group activities, a third of the school day in free-choice activities or interest centers, and the remaining third in routines and meals (Early et al, 2010; Cabell et al, 2013; Pianta et al, 2018; Nores et al, 2022). On average, classrooms spend a larger proportion of time in arts, language, and literacy activities than in math and science (Cabell et al., 2013; Connor et al., 2009; Early et al., 2010; Justice et al., 2021; Pianta et al., 2018).

There is no clear pattern in how much time children spend in small groups and individual instruction, potentially due to differences in how measures of these two learning settings are operationalized across studies. For example, three studies using the same measurement instrument (Emerging Academic Snapshot, EAS) report time spent in learning settings as mutually exclusive (Cabell et al., 2013; Chien et al., 2010; Early et al., 2010) whereas one study requires double coding instances of recess/outside time in conjunction with another appropriate learning setting (Fuligni et al., 2012). The latter study found that children spent approximately 10 percentage points (pp) more time in free choice/centers, and twice or thrice the proportion of time in small group settings when compared to the former studies. However, it is not possible to assert whether these increases in time spent in free choice/centers and small groups represent different time use patterns or are due to the inclusion of recess time as a double-coded learning setting. Similarly, some studies of time use do not report time children spend in individual time (Bratsch-Hines et al., 2019; Burchinal et al., 2021), and another study on this topic includes an additional learning setting —dyads—not explored in other work (Justice et al., 2021) (see Table 1.1).

Findings about content domains across these studies show that prekindergarten classrooms emphasize language and literacy instruction more so than instruction in other developmental domains, with significant variation—ranging from 14% to 46% of time spent in language and literacy—across studies. Studies also suggest that children's time spent in science, social studies, art, and math differs across samples, with social studies and art each accounting for approximately 15% of the school day (see Table 1.1) and math, science, and motor development taught to a lesser extent. Again, making conclusions about time use across these studies is challenging due to the measurement differences in time-sampling strategies, authors' choices about whether content areas are mutually exclusive, and what areas are accounted for with each instrument. For example, studies in this review vary on the number of observation days (ranging from one to four), whether the observation period includes only mornings or extends to a full-day, and whether routines are included within content areas or activity settings (see Table 1.1). Similarly, some study protocols require that specific content areas be observed, such as literacy or math instruction (Bratsch-Hines et al., 2019; Burchinal et al., 2021; Weiland et al., 2023), whereas other studies do not report aiming to capture specific content (Cabell et al., 2013; Chien et al., 2010; Early et al., 2010; Fuligni et al., 2012, 2012; Justice et al., 2021; Nores et al., 2022; Pianta et al., 2018). In sum, prior studies have identified variation in classroom time use but measurement differences do not allow us to consistently identify patterns representative of classrooms' instructional features. We address this gap by examining descriptive consistency across measures. Further, we examine whether such consistency also interferes with conclusions when using a person-centered approach to identify classrooms' time use profiles.

Time Use Profiles (Person-Centered Approach)

Learning formats and content areas are closely linked components of instruction that are observed in tandem in real world settings. To address this complexity, researchers have also described multiple features of classrooms by creating time use *profiles*, or systematic variation across classrooms explained by combinations of their instructional characteristics (Chien et al., 2010; Fuligni et al., 2012; Justice et al., 2021). As shown in Table 1.2, at least three studies have identified either two or four instructional profiles based on time distribution across learning settings and activities. In studies identifying two profiles,[1] classrooms differed in the proportion of time spent in free-choice settings (Fuligni et al., 2012) and in the proportion of time spent in whole group working on language, literacy, and math activities (Justice et al., 2021). Children spent less than 15% of their time receiving direct instruction in whole- or small-group arrangements in a free-choice profile (Fuligni et al., 2012) and around 20% of time working in whole group settings in an academic-light work profile (Justice et al., 2021). One study identified classrooms that used the majority of time in a) whole-group instruction and language and literacy instruction, b) whole group instruction but distributing time across varied content areas, c) individual instruction, and d) free choice or small group instruction, with less emphasis on content areas.

Time Use Variation Across Demographic Groups

There is some evidence that time use profiles vary with respect to the demographic makeup of classrooms (Chien et al., 2010). For example, children from families with low incomes attending classrooms with a larger proportion of time spent in individual instruction showed larger gains than their higher-income peers in the same profile. In all other classroom profiles, children from families with low incomes had smaller gains than their peers. Research using variable-centered approaches suggests similar patterns. For example, Early et al. (2010) showed that classrooms with a higher proportion of Latino and Black children spent less time in free-choice activities (*d* = -0.53 and -0.76, respectively). Similarly, classrooms with a higher

[1] Justice et al. (2021) identified four classroom profiles from prekindergarten to third grade, but prekindergarten classrooms only fit in two profiles. Although we present their full set of profiles in Table 1.2, comparisons are restricted to groups effectively including prekindergarten classrooms.

proportion of Black children spent more time in meals and routines (*d* = .36). Identifying potential demographic and quality differences in classroom profiles can inform efforts to reduce opportunity gaps by kindergarten entry, especially if there are benefits or drawbacks of specific profiles for children from marginalized groups.

Associations Between Time Use and Children's Academic Gains

Scholars have argued that understanding time use can also address the current need in the field to develop measures of prekindergarten quality that are linked to gains in children's academic and cognitive skills. Studies in this area have found small associations between time spent in instructional learning formats and gains in children's skills. For example, in a sample of 63 classrooms in rural North Carolina, Bratsch-Hines and colleagues (2019) found that children exhibited larger gains in reading decoding (as measured by Woodcock Johnson Letter-Word Identification, *g* = 0.13) and phonemic awareness (as measured by Dynamic Indicators of Basic Early Literacy Skills —Phonemic Segmentation Fluency, *g* = 0.15) when they spent more time in small groups. Conversely, spending more time in large-group settings was negatively associated with gains in children's expressive language skills (as measured by the Expressive One-Word Picture Vocabulary Test, EOW, g = - 0.12). Burchinal and colleagues (2021) replicated some of these associations with the same North Carolina prekindergarten classrooms when accounting for the Classroom Assessment Scoring System (CLASS PreK; Pianta et al., 2008) domains in their models, demonstrating that associations were robust even after accounting for the general quality of interactions in the classroom. Relatedly, Burchinal et al. (2021) found that more time spent in whole groups negatively predicted gains in children's math skills.

Although fewer studies have used the proportion of time-spent in *content areas* to predict children's academic gains, there is some recent evidence to show that they may also matter. Bratsch-Hines and colleagues (2019) found null associations between time spent in learning activities related to oral language and literacy and gains in children's expressive language and reading (decoding), respectively. However, they found a positive association between the proportion of time in letter-sound activities and DIBELS First Sound Fluency scores (g – 0.13). Burchinal and colleagues (2021) also found associations between the proportion of time spent in literacy activities and children's gains in expressive language skills (SE = 0.08); and the proportion of time spent in sound focused activities and children's phonemic awareness (SE = 0.08) and first letter recognition (SE = 0.08). In a study conducted in the BPS prekindergarten program, there was also a statistically significant association between time spent in language/literacy and children's executive function gains (Weiland et al., 2023). In sum, time use measures have significantly predicted children's gains in language, literacy, math, and executive function skills in prekindergarten classrooms in prior work, although not consistently across studies.

Interestingly, time spent in "free play" or being in a classroom with "light academic profile" appears to be associated with smaller gains in literacy and math skills, in comparison with other instructional profiles such as those with more time dedicated to whole-group or individual instruction (Chien et al., 2010; Fuligni et al., 2012; Justice et al., 2021). Children spending the bulk of time in free-play experienced the smallest literacy gains in skills such as naming letters (partial η^2 = 0.03) and letter-word identification (partial η^2 = 0.04). Similarly, children in the same group showed smaller math gains on the Woodcock Johnson Applied Problems subtest (partial η^2 = 0.01), naming numbers (partial η^2 = 0.01) and counting (partial η^2

= 0.02), compared to children attending classrooms privileging individual instruction, whole group instruction, and scaffolded instruction. The individual instruction profile outperformed all other groups on the Woodcock Johnson Applied Problems subtest (Chien et al., 2010).

Children attending “Structured-Balanced” classrooms—namely, classrooms with similar proportion of their days engaged in distinct instructional formats and spending similar amounts of time across content areas (e.g., literacy, math, art, and arts)—showed larger gains in language skills (*SD* = 0.35) compared to children attending classrooms in the “High Free-Choice” profile (i.e., spending on average 61% of the day in free-choice activity settings). However, there were no statistically significant differences in children’s math or self-regulation gains, based on profiles drawn from a sample of center-, family-, and school-based settings (Fuligni et al., 2012).

In sum, there is evidence that children in classrooms dedicating more time to instructional activities (i.e., with specific learning purposes) have larger developmental gains in academic skills than children in classrooms dedicating more time to non-instructional activities. However, profiles inconsistently predicted gains in math. These results may reflect the fact that prekindergarten classrooms generally spend a small amount of time in math instruction. The Boston prekindergarten program that is the focus of our current study is a potential exception, with researchers finding that children spend about 36 minutes per day in math instruction (Weiland et al., 2023), possibly because these classrooms implement an evidence-based math curriculum as part of typical practice. There is a clear need to develop a stronger understanding of how best to collect and apply time use measures not only to describe what is happening in classrooms, but also to determine whether these tools can predict child gains and be used to make actionable decisions to strengthen classroom quality and students’ outcomes.

Present Study

The early education field has yet to identify whether the variation in time use that has been described across several studies and samples is due to differences in measurement instruments or data collection procedures or to real differences across settings. Our study builds on the prior prekindergarten time use literature by holding constant several important sources of variation in observed time use—curricula used in the setting, content taught during the observation period, length of observation, and characteristics of children—to compare prekindergarten time use across two different measures. In doing so, we describe to what extent these measures—the Individualizing Student Instruction observation (Connor et al., 2009) and the Narrative Record (Farran et al., 2015)—capture similar classroom time use profiles in a sample collected in the Boston Public Schools prekindergarten program. We then describe different arrangements of instructional content areas and formats throughout the day, namely instructional profiles, and explore whether they are predictive of children's academic gains.

Our specific research questions are:

1. Do the NR and ISI produce similar time use descriptions of BPS Prekindergarten classrooms?
2. How do time use profiles based on the NR and the ISI compare in terms of the number of profiles derived by each instrument, their description, and potential differences in children and parent demographics, classroom quality, and classroom composition?
3. Do time use profiles of instructional content and formats based on the NR and the ISI predict children's gains in language, math, and working memory, above and beyond differences in profile membership such as classroom quality and demographic composition?

Method

Participants and Setting

The sample consists of 247 students (52% female, age = 4.66 years; *SD* = 0.29) enrolled in 35 prekindergarten classrooms in 20 schools implementing the BPS prekindergarten curriculum and professional development model during the 2016–2017 school year. On average, 49% of students in the sample were Dual Language Learners (DLL), 48% were eligible for free or reduced-price lunch, 46% were Hispanic, 26% were Black, 16% were White, 9% were Asian, and 3% were mixed race or another race. About 40% of third-grade students in study schools met or exceeded expectations on the 2015–2016 state English/Language Arts exam, and 45% met or exceeded expectations on the state math exam. Although the schools in the sample are generally representative of the population of BPS elementary schools offering a prekindergarten program, schools in our sample had lower proportions of Black students (32% at the district level) and higher proportions of students meeting or exceeding expectations on the 2015–2016 ELA exam (36% at the district level).

In public school settings, the BPS prekindergarten program is free, runs full-day, and is open to any age-eligible child for the academic year. All BPS prekindergarten teachers meet the same requirements and receive the same compensation as K-12 teachers and are required to have an early childhood (preschool to grade 2) license from the Massachusetts Department of Elementary and Secondary Education and have or be working towards a master's degree in education. All classrooms included in the current study implemented the BPS Focus on K1 curriculum, with implementation supported through district-provided training and coaching. Teachers in the sample had on average 9.69 (*SD* = 7.66) years of experience teaching prekindergarten and 80% of them had a master's degree.

Procedures

The Institutional Review Boards at the lead organization for this study approved the human subjects plan prior to the commencement of study activities. The project name is ExCEL P-3: Promoting Sustained Gains from Preschool to Third grade and the study was approved by the MDRC IRB (approval number 860661-2).

School and Classroom Recruitment

The research team randomly selected 25 schools from the full set of 76 schools that offered the public prekindergarten program in 2016-2017. Four schools declined to participate, and one was designated as a pilot school for developing new measures and was excluded from the study. All prekindergarten teachers assigned to general education or inclusion classrooms in each of the 20 participating sample schools were invited to participate in the study in the fall of 2016. Ninety-six percent (N = 41) agreed to the study activities. However, two prekindergarten teachers in one school declined to be videotaped. The team was able to collect two observations and code videotapes on both NR and ISI for 35 classrooms given existing resources. The four classrooms excluded did not differ systematically from the classrooms retained, resulting in an analytical sample of 35 classrooms in 17 schools.

Classroom Videotaped Observations

During the Winter of 2017, each classroom was videotaped for two hours during two visits scheduled in advance with teachers. In addition to the lead teacher, a paraprofessional was present on visit days in 88% of the classrooms. The research team 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 starting coding, we synchronized videos from the two observations to effectively track both the teacher and students as they moved between camera angles. Two independent teams coded the study videos using the NR and the ISI protocols. The ISI was coded using both camera angles and the NR was coded using the lead teacher videotapes.

To score classrooms on the Narrative Record, coders participated in a one-day training and had to pass a reliability test, demonstrating a minimum of 80% agreement with a master coder, prior to coding. Then, coders used a predetermined format in Microsoft Excel – which built off the version used by the developers of the measure (Farran et al., 2015) to record the duration of activities as indicated by the measure tool. We randomly selected 25% of classrooms to double code throughout the coding process. Following prior work (Farran & Bilbre, 2014), we estimated inter-rater agreement across activity settings and content categories, within 3% of the total minutes observed in each episode. Coders' average agreement was 81% (Activity type = 77%; Content category = 86%). This set of procedures mapped onto those used in a similar large-scale study of prekindergarten conducted in New York City (Morris et al., 2016). To score classrooms on the ISI, coders participated in multiple training sessions on ISI measures and were tested on the mastery of the codebook before coding. We used Noldus Observer XT 13 software for coding videotapes (Noldus Information Technology, 2013). After training, coders had to show reliability on the ISI via coding four 20-minute video segments. Compared to a master-coded file, all coders scored >.80 Kappa on each of the four videos. We randomly selected and double-coded 20% of the video observations throughout the coding process to prevent drift in inter-rater reliability. 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-second grace window. Our average Kappa was 0.76, similar to past ISI studies (e.g., average of .76 in Connor et al., 2009).

Additionally, the team collected measures of classrooms' process quality from the same video recordings using the Classroom Assessment Scoring System (CLASS; Pianta et al., 2008). For the CLASS (Pianta et al., 2008), coders participated in a two-day training and then established reliability on a set of master codes created by the CLASS developers. As recommended by the measure's protocol (Pianta et al., 2008), coders used cycles of 20 minutes for observing and 10 minutes for scoring, which they repeated four times for each observation. Coding began when instruction commenced in the video and ceased after 80 minutes of observed time. We double-coded 20% of the observations to assess interrater reliability. Throughout the coding process, we conducted drift checks wherein observers had to code a master tape every three weeks to ensure they stayed reliable across time. 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.

Student Recruitment and Direct Assessments

From late September 2016 through late November 2016, the research team solicited informed consent from all students in participating classrooms. Overall, 81% of children in participating classrooms agreed to participate. The research team randomly selected 50% (~6–10 per classroom) of consented children to participate in student-level data collection activities for 263 prekindergarten students, from which 263 (86%) attended classrooms that were videotaped. We trained research staff to reliability and then collected direct assessments of academic skills in the fall of 2016 and spring of 2017. We used the Pre-language Assessment Scale (preLAS; Duncan & DeAvila, 1998) Simon Says and Art Show tests to determine the administration language for a subset of assessments. The preLAS assesses preliteracy skills and proficiency in English. Of the 247 children in the current study sample, 31 did not pass the preLAS and completed a subset of assessments in Spanish in the fall and 12 students did not pass the preLAS and completed assessments in Spanish in the spring.

Parent Surveys

We used text messages and emails to contact the consenting parents of all students who were selected for the study sample and collected parental demographic information via 20-minute surveys in the fall of prekindergarten (2016). The research team sent biweekly reminders to complete the survey, and paper copies were sent home to non-responding parents via backpack mail. All parents received a $25 gift card for completing the survey. Parents completed the survey for 93% of the 247 children in the current study sample.

Teachers Survey

In the spring of prekindergarten, we asked teachers to complete a survey reporting on their demographic characteristics, teaching experience, and instruction. We used data on demographics and teaching experiences to describe the sample. Out of 35 teachers in the analytic sample, 100% responded to the survey.

Measures

Classroom-Level Measures of Time Use

Narrative Record (NR). The NR (Farran et al., 2015) is an open-ended format that allows observers to describe the duration of events in the classroom for episodes of time. Episodes are based on whether children are engaged with an activity type (i.e., whole group, small group, centers, a small group/centers —indicating when some children are in small groups and other children are in centers, simultaneously; and transitions) and with a particular domain of instructional content (i.e., language/literacy, math, social studies, science, motor development activities, socio-emotional learning), or a combination of content areas which is coded separately as mixed. The NR also includes a “no content” code, used when no activity with a recognizable instructional purpose is occurring (e.g., during transitions, meals, and nap time). Each episode begins when at least 75% of children engage with a new activity type (for the remainder of this paper we will refer to these as “learning settings” to facilitate comparisons across instruments) or a new activity and finishes when most children switch to a new learning setting, activity, or a combination of both. Importantly, the NR protocol states that when an ongoing activity is interrupted for less than a minute, a principle of continuity prevails, and therefore the episode continues. However, if the interruption lasts at least a minute, a new episode is added and scored as a transition. Following the authors’ protocol (Farran et al., 2015), we estimated the total duration of each episode in minutes. We added a classroom’s data together across both observation days to create measures for each NR construct of interest and then obtained the percentage of time spent in each learning setting and activity.

Individualizing Student Instruction (ISI). We used the Individualizing Student Instruction (ISI) Coding System (Connor et al., 2009) to obtain child-level measures of time use for a selection of children (3 to 10 per classroom) that were later aggregated to the classroom level. The ISI captures continuous measures of *quantity of time* (e.g., 0 minutes – 58 minutes) an individual child is engaged in different learning settings (i.e., whole group, small group, centers, and individual learning); in distinct instructional or non-instructional activities (i.e., language/literacy or math versus time spent in transitions); and in different content domains (e.g., social studies, science, motor development activities, socio-emotional learning), throughout the full duration of an observation. Importantly, the ISI coding system allows coders to enter up to two main content domains involved in activities, so the amount of time when children are learning integrated content areas is obtained through data processing (i.e., estimating the proportion of time when two different content domains are taught simultaneously). This characteristic of the ISI allows researchers to retain the specificity of instructional content while indicating that several domains were taught in conjunction. Each classroom observation was coded following an adaptation of the protocol designed by Connor et al. (2009) for early childhood education settings (see Weiland et al., 2023). We coded each second of observed time for each child, switching codes as necessary to capture children's settings, activities, and content domain. For the students with two observations (84%), we first summed their data across both observation days to create aggregate child-level measures for each ISI construct of interest. We then obtained the total number of minutes for each specific code at the child level. To facilitate comparisons between the ISI and the NR (recall the latter is coded at the classroom level when 75% or more of the children are engaged with a given activity), we calculated the percentage of time spent in each learning setting and instructional activity and aggregated ISI measures at the classroom level.

Classroom Process Quality and Indicators of Structural Quality. We coded general 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 domains are scored on a 7-point scale. In prior work with our prekindergarten sample, the CLASS did not predict gains in children's outcomes (Guerrero-Rosada et al., 2021).

Measures of teachers' experience and education were constructed based on survey data. Teachers reported their highest level of education, from which we created an indicator of whether the teacher holds a masters' degree. They also reported the years of experience teaching in prekindergarten, which we used as a continuous measure.

Parent Characteristics

We constructed indicators of the reporting parent's level of educational attainment as a proxy for socio-economic status (high-school, two-years degree, bachelor's degree, graduate degree); whether there was at least one parent in home working full-time (35 hours/week or more); and whether the parent was married or lived with a partner. We also used continuous measures to describe the age of the child's mother at her first birth, the parent respondent's age in the fall of 2016, and the number of people living in the household. We include these to match prior work with this sample (Guerrero-Rosada et al., 2021; McCormick et al., 2020, 2021; Weiland et al., 2023); experts have advised including the same covariates across studies from the same dataset to prevent illusionary results (Gehlbach & Robinson, 2017).

Children's Demographic Characteristics and Classroom's Composition

We accessed administrative records from the school district on children's demographic characteristics to create indicators of their gender; eligibility for free or reduced lunch; dual language learner status (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); race (Asian or Asian American, Black, other or mixed race, and White); ethnicity (Latinx/Hispanic); and birthdate, which we used to calculate children's ages when their baseline measures were collected. We also used administrative records to create measures of classrooms' demographic composition for the same groups described above based on the population of enrolled children in the 2016 – 2017 school year.

Direct Measures of Children's Language, Math, and Working Memory Skills

Field-based data collectors assessed children’s receptive language skills in the fall and spring of the prekindergarten year using the Peabody Picture Vocabulary IV (PPVT IV; Dunn & Dunn, 2007). The PPVT IV is a nationally normed measure that has been used widely in diverse samples of young children, has excellent split-half and test–retest reliability estimates, and strong validity properties (Dunn & Dunn, 2007). It requires children to choose which of four pictures best represents a stimulus word, verbally or non-verbally. We used the raw score total as our outcome measure in our primary analysis, and present models using the age-standardized scores in Appendix A. The research team assessed all children on the PPVT—regardless of whether they passed the PreLAS language screener—to describe an equivalent measure of receptive language skills in English across the full sample.

We assessed children’s math skills using the Woodcock Johnson III Applied Problems subtest (WJ-AP III; Woodcock et al., 2001, 2005) and the Research-based Early Mathematics Assessment (REMA; Clements & Sarama, 2008). The Woodcock–Johnson Applied Problems subtest requires children to perform relatively simple calculations and solve arithmetic problems. Its estimated test–retest reliability for 2- to7-year-old children is 0.90 (Woodcock et al., 2001) and it has been nationally normed and used with diverse populations of children (Gormley, Gayer, Phillips, & Dawson, 2005; Wong, Cook, Barnett, & Jung, 2008). The research team assessed Spanish-speaking children who did not pass the PreLAS language screener using Batería III Woodcock Muñoz (Woodcock, Munoz-Sandoval, Ruef, & Alvarado, 2005), which follows similar norms to the Woodcock–Johnson English version and allows for combining scores across both English and Spanish in the sample. We used raw scores for our main models, but we present results using the age-standardized version of the test in Appendix A. In our sample, the majority of children completed the test in English (6.5% of the sample completed the assessment in Spanish in the fall and 2% completed it in Spanish in the spring).

We also measured math skills using the REMA (Clements & Sarama, 2011), a hands-on assessment of children's early math skills (e.g., numeracy, geometry, operations, spatial reasoning). The alpha reliability of the test subscales ranges from r = 0.71 (geometry) to 0.89 (numeracy). We present results using the REMA raw score in our main analyses. We did not assess children on the REMA during the fall and thus use the Woodcock–Johnson Applied Problems score as a baseline for all models examining math skills.

We assessed children's working memory using the Digit Span Forward test (DSF; Rosenthal et al., 2006), which 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. This test is widely used and nationally normed. We used the categorical score for Forward Digit Span (FDS), representing the sequence with the highest number of digits the child repeated accurately. This test has high correlations with Backward Digit Span and other executive function tasks and has shown good test–retest reliability in samples of prekindergarten children (r = 0.80; Muller, Kerns, & Konkin, 2012).

Analytical Approach

To identify *whether the NR and ISI produce similar time use descriptions of prekindergarten classrooms* (RQ 1), we used t-tests to compare the proportion of time children were engaged with each type of instructional domain (e.g., language, math, social studies) and learning formats (e.g., small group, whole group), as measured by the NR (classroom-level) and the ISI (child-level, aggregated to the classroom level).

Then, to examine *how time use profiles based on the NR and the ISI compare* (RQ2), we used latent profile analysis (LPA) to identify time use profiles for each instrument based on the proportion of time classrooms spent in learning settings and activities. LPA aims to identify latent homogeneous subgroups within a heterogeneous sample based on a certain set of continuous variables (Rosenberg et al., 2018; Spurk et al., 2020). Prior work in the field has shown that groups of classrooms can be defined by their instructional characteristics, for example, those using a larger proportion of time working on language / literacy activities while in whole-group formats, or those using a larger proportion of time in free-choice activities (Fuligni et al., 2012; Justice et al., 2021). As mentioned in our literature review, prior work in the field has found that prekindergarten classrooms group into four distinct instructional profiles: a free-play or academic-light group, a whole-class academic focused group, a group that balances content and formats of instruction, and a group that privileges individualized work (Chien et al., 2010; Fuligni et al., 2012, see more in Table 2). Thus, we test the hypotheses that classrooms are best described by two, three, or four profiles defined by how teachers distribute time across learning settings (i.e., whole group, centers, small group, and individual time) and instructional content (i.e., math, language, other content areas, and mixed content), against the null hypothesis that classrooms are best described by a single profile. Following best practices in LPA, we use a combination of empirical and conceptual criteria to determine the best fitting model for each instrument (Spurk et al., 2020). We then regressed classrooms' profile membership on a combination of classroom characteristics: process quality scores and indicators of structural quality; children's baseline scores; parents' characteristics; and classrooms demographic composition for each instrument. We did so to evaluate whether classrooms differed in factors other than their learning settings and instructional content by profile membership. We chose equivalent or comparable profiles across instruments as the reference group.

Finally, we sought to describe whether time use profiles of instructional content and formats based on the NR and the ISI predict children's gains in language, math, and working memory, above and beyond demographic differences in profile membership. We first estimated multi-level regression models predicting children's spring scores in vocabulary, numeracy, and working memory skills, accounting for the baseline level of the outcome or proxy (in the case of the spring REMA score), and additional sets of child-, parent-, and classroom-level covariates, following similar procedures used with these data by Maier and colleagues (2022) (1):

$$Outcome_{ijk} = \beta_0 + \beta_1 Classroom\ Profile_{jk} + \beta_2 Baseline_{ijk} + \chi_{ijk} + \rho_{ijk} + \lambda_{jk} + \mu_k + \gamma_{jk} + \varepsilon_{ijk} \tag{1}$$

where the subscript *i* refers to an individual student, *j* denotes an individual classroom, and *k* represents an individual school. $Outcome_{ijk}$ refers to children's spring scores in vocabulary and numeracy. The key predictors are a set of indicators of membership to each classroom profile $\beta_1 Classroom\ Profile_{jk}$ —using the profile with the most commonalities across the NR and the ISI as the group of reference. All models control for the children's corresponding baseline score $\beta_2 Baseline_{ijk}$. χ_{ijk} is a vector of student-level characteristics including their race/ethnicity, gender, DLL status, eligibility for free or reduced lunch, age, and testing interval from Fall to Spring. $\rho_{ijk}$is a vector of parent characteristics measured at the child level including indicators for whether the parent works, is married, completed a two-year degree, completed a bachelor's degree, or completed a graduate degree (with parents who completed high school as the reference group), responding parent's age, mother's age at first birth, and the number of people in the home. λ_{jk} is a set of characteristics measured at the classroom level that includes the CLASS Instructional Support score, teacher experience, and an indicator of whether the teacher has a master's degree, which we included to disentangle profiles based on the quantity of instruction (i.e., classroom instructional profile) from process and indicators of structural quality. Models include random intercepts for school and classrooms, μ_k, γ_{jk} and ε_{ijk} are the school-, classroom-, and student-level residual terms.

As a second step, following Zamarro and colleagues (2015), we added a set of classroom composition measures (% students eligible for free or reduced lunch, % girls, % dual-language learners, % Black, % Hispanic, % Asian, % Other, with % White as the reference group) to examine whether associations between classroom's instructional profile and children's gains remain stable after controlling for differences in the profiles that could be associated with classroom's demographic composition.

Results

Descriptive Statistics

Our sample was racially, linguistically, and socio-economically diverse (see child-level characteristics in Table 1.3). CLASS scores ranged from moderate to high in Emotional Support and Classroom Organization, and from low to moderate in Instructional Support. There was high variability in classrooms' socio-economic and demographic composition (e.g., ranges of % Asian = 0–100, % Black = 0–88, % Latino = 0–88, % Other = 0–21, % White = 0–70, % DLL 0–100, and % FRPL = 8–100). We explore this variation further in relation to classrooms' instructional profiles.

RQ 1: Do the NR and ISI Produce Similar Time Use Descriptions of Prekindergarten Classrooms?

The NR and ISI partially produced similar time use descriptions of prekindergarten classrooms. For learning settings, both instruments showed similar results regarding the proportion of time children spent in whole group and transitions (see Table 1.4). However, the proportion of time children spent in small groups and centers was statistically significantly different by instrument, with a higher proportion of time in centers and small groups with the ISI than the NR. These differences likely reflect ISI's focus on capturing children's *individualized* instruction and the NR's focus on capturing the overall group experience (i.e., using a code to identify when children are simultaneously working in centers and small group settings).

For learning activities, the two instruments were consistent in capturing the proportion of time that children were exposed to math, science, social studies, socio-emotional, and motor development activities (see Table 1.4). They were also consistent in measuring the proportion of time when content areas were combined in the same learning activities (i.e., defined as "mixed content" by the NR and as "integrated content" by ISI). There were statistically significant differences in the proportion of time dedicated to language/literacy and arts, with ISI capturing a higher proportion (8 pp, $p < 0.001$; and 3 pp, $p < 0.01$; respectively) of time spent in both areas than the NR. There was also a statistically significant difference in the proportion of time children were not engaged with learning activities, with the NR capturing a higher proportion (7 pp; $p < 0.05$) of non-instructional time than ISI.

RQ 2. How do Time Use Profiles Based on the NR and the ISI Compare?

We conducted LPA for each instrument and used three comparison criteria to examine consistency of results across instruments: a) total number of profiles derived by each instrument, b) substantive descriptions of profiles, and c) contextual differences by profile membership.

Total Number of Profiles by Instrument

Classrooms' time use as measured by the NR was better represented by two, three, and four profiles, compared to the one profile solution, as indicated by statistically significant changes in the Bootstrapped Likelihood Ratio Test (BLRT) and increases in log likelihood (see Table 1.5). The two-profile solution differentiates three classrooms from the rest in the sample. However, this is less substantive informative than the subsequent three-profile solution, which retains the original small group and allows for a better differentiation among instructional modes in the remaining classrooms. Although a four-profile solution showed better entropy (i.e., confidence with which classrooms were assigned membership to a given class), SABIC and AIC statistics are lower for the three-profile solution. Additionally, two small groups containing three classrooms would create statistical power challenges with implications for comparison and predictive models.

A four-profile solution is statistically preferred for the ISI compared to other solutions, as indicated by statistically significant changes in BLRT and increases in log likelihood (see Table 1.5). However, a four-profile solution is not conceptually informative since it generates a profile with a single classroom. A three-profile solution (with profiles described in more detail below) is conceptually sound, shows adequate entropy, has a smaller AIC and SABIC than the four-profile solution, and is comparable with the preferred solution resultant from the NR measures.

Profiles Substantive Descriptions

In sum, we found a profile focused on language/literacy instruction primarily using whole group instruction, a profile dedicating most of the time to integrated or mixed content while using varied learning formats, and one additional profile for each instrument including a small number of classrooms dedicating time to other content during whole group according to the ISI or centers and small groups simultaneously according to the NR. Regarding profile descriptions (see Table 1.6), LPA results show a *Whole Group / High Academic (WGA)* profile, similar across instruments in its time distribution for learning settings and content areas. This profile was predominant in the NR (54% of classrooms, vs. 40% in ISI), with an overlap of 10 classrooms across instruments (29% of the total sample, 48% of the profile resultant from the NR, and 71% of the profile resultant from the ISI). *WGA* classrooms spent half the time learning in whole group settings. Children in NR *WGA* classrooms were engaged in language instruction 25% of the time, almost ten percentage points more time than children in other classrooms were. This represents 41 minutes of language and literacy instruction, compared with 18 and 21 minutes in the SGC and Balanced Mixed-Content profiles, respectively. Regarding math, children in the NR *WGA* profile engaged with math activities 15% of the time (or the equivalent of 23 minutes), in comparison to 6% (7 minutes) and 11% (16 minutes) in the SGC and High-Transitions profiles, respectively (see Appendix B).

Similar to what we observed for the NR *WGA*, children in ISI *WGA* classrooms were engaged in language instruction 34% of the time, almost ten percentage points more time than children in other classrooms were. This represents 70 minutes of language and literacy instruction, in comparison with ~42 minutes in other profiles. Regarding math, children in the ISI *WGA* profile engaged in a similar proportion of time in math activities in comparison with other profiles. Classrooms assigned to this profile for the ISI were coded for a higher number of minutes (N = 14, Mean = 211 minutes) than classrooms assigned to this profile for the NR (N = 19, Mean = 157 minutes, $p < .05$) despite using the same videotapes. This finding suggests that classrooms in this profile had larger segments between episodes as defined by the NR measure.

Results also suggest a second overlapping profile: *Balanced Mixed (BM)* classrooms were similar across instruments in time distribution across learning settings and activities, except for time spent in language/literacy instruction. This profile is predominant in the ISI (46% of classrooms, vs 37% in NR), with an overlap of 9 classrooms across instruments (26% of the total sample). Regarding learning settings, children in *BM* classrooms spent a similar proportion of time in whole group, centers, transitions for both instruments, and individual time for the ISI profile. Regarding learning activities, children in NR *BM* classrooms spent a similar proportion of time learning language/literacy and math content (14% and 11% respectively, equivalent to 21 and 16 minutes), but dedicated most of the time to learning a combination of content areas (39%, equivalent to 63 minutes) across observation days. Children in ISI *BM* classrooms spent twice the proportion of time in language/literacy activities in comparison to math and other content activities (24%, 12%, and 11% respectively). ISI *BM* classrooms spent 41% of time learning mixed content (equivalent to 71 minutes), which is similar to the 39% as observed with the NR (equivalent to 63 minutes). Given that the ISI coding system retains the specific content areas when children engage with integrated activities—as explained in the measures section—we note children in *BM* profiles primarily engage with language/literacy activities presented in combination with math and other content areas.

The third NR and ISI profiles were not equivalent, although they were similar in their small size. Based on the NR measures, there were three classrooms primarily learning in a combination of Small Groups and Centers (*SGC)* settings (54%, equivalent to 61 minutes). Children in these classrooms also spent most of the time (62%, equivalent to 71 minutes) in activities that combine content areas. Classrooms in the NR *SGC* profile spent less time in transitions compared with other profiles (14% of time, equivalent to 16 minutes; vs. 19% and 30% of time, equivalent to 30 and 42 minutes respectively). Based on the ISI, there were five classrooms spending most of the time in *Whole Group* settings (64%, equivalent to 98 minutes) integrating content areas (*WGM*). Although classrooms placed in the *WGM* profile by the ISI dedicated 7 pp less time to language/literacy instruction compared to *WGA*¸ children in these classrooms doubled the proportion of time (19% vs. 8%) they spent in other content activities (e.g., social studies, sciences, socio-emotional learning). The proportion of time children in *WGM* spent in math was similar to the proportion of time in other ISI profiles. Children in this profile engaged with activities combining content areas 64% of time. Due to the small number of classrooms in these two profiles, we exclude them from subsequent analyses.

Contextual Differences by Profile Membership

Next, we estimated a regression model for each instrument predicting classrooms' profile membership based on a combination of characteristics such as their process and structural quality, children's baseline, parent's demographics, and classroom composition. Results are shown in Table 1.7. We discuss to what extent profiles are statistically significantly different in aspects other than their instructional characteristics, by instrument, below.

Differences in Classroom Quality. Instructional profiles had differences in their process and structural quality, although the directions of such differences were inverse across instruments. Classrooms in the NR *BM* profile had 0.77 *SD* ($p < 0.01$) higher Emotional Support and 0.86 *SD* ($p < 0.001$) lower Instructional Support than classrooms in the NR *WGA* profile. Conversely, classrooms in the ISI *BM* profile had 0.82 *SD* ($p < 0.001$) lower Emotional Support and 1.03 *SD* ($p < 0.001$) higher Instructional Support than classrooms in the ISI *WGA* profile. There were no statistically significant differences between profiles in Classroom Organization scores for any of the measures. Regarding structural quality, classrooms in the NR and ISI BM profiles were part of schools in which average class size was -0.02 *SD* ($p < 0.05$) and -0.03 *SD* ($p < 0.001$) smaller than classrooms in *WGA* profiles. Teachers of classrooms in the ISI *BM* profile were 0.33pp more likely to have a Master's degree (*p* < 0.001) than teachers in classrooms with the ISI *WGA* profile.

Differences in Children's Baseline Characteristics. There were no statistically significant differences by classroom profile for either measure, in children's language, math, working memory skills, or age. The exception is a statistically significant coefficient differentiating children's math baseline scores for the ISI profiles, but the size of this difference is zero (-.00 *SD*, *p* < 0.01).

Differences in Parents' Characteristics. The proportion of children living with married caregivers was the only statistically significant difference between both NR and ISI *BM* classrooms and *WGA* classrooms (-0.33 and -0.25 SD, *p* < 0.05, respectively). There is also a higher proportion of working caregivers in the ISI *BM* group than the ISI *WGA* group (0.46 *SD*, *p* < 0.05).

Differences in Classroom Composition. Classroom composition was statistically different across profiles. In particular, NR *BM* classrooms had a larger proportion of children eligible for free or reduced priced lunch (7 pp, *p* < 0.001) and a larger proportion of dual-language learners (12 pp, *p* < 0.05) compared to NR *WGA*. When holding constant the proportion of White children in the classroom, NR *BM* classrooms had fewer Latino (15 pp, *p* < 0.001) and Asian (3 pp, *p* < 0.001) children than classrooms in the *WGA* profile. Conversely, ISI *BM* classrooms had a smaller proportion of children eligible for free or reduced lunch (11 pp, *p* < 0.001), a larger proportion of dual language learners (5 pp, *p* < 0.001), lower proportion of Latino children (4 pp, *p* < 0.001), and a higher proportion of Black children (14 pp, *p* < 0.001) and children from two or more races (3 pp, *p* < 0.001) compared to ISI *WGA*.

RQ 3. Do Time Use Profiles Predict Children's Gains, Above and Beyond Classroom Process Quality and Demographic Differences in Profile Membership?

As shown in Table 8, there was no evidence of predictive validity for the profiles derived from these measures in relation to children's language and working memory gains. There were differential math gains by profile membership, but these depended on the outcome measure used, and were not robust to controls for classroom quality and classrooms' demographic composition. Children attending classrooms in the *NR WGA* profile had larger math gains on the Woodcock Johnson Applied Problems subtest than children attending *NR BM* classrooms (*standardized association* = 0.23, $p < 0.05$), even after controlling by classrooms' process and structural quality (*standardized association* = 0.27, $p < 0.05$), but this association was no longer statistically significant once models accounted for classrooms' demographic composition. For the ISI measures, profile membership was not a statistically significant predictor of gains in any of the examined skills, except for a small association with math when measured with the REMA (*standardized association* = 0.21, $p < 0.05$). This association was no longer statistically significant once we accounted for classroom quality and demographic composition. Results were consistent in models using standardized outcome measures (see Appendix A).

In this paper, we aimed to identify the extent to which time use measures provide reliable and useful information about classroom instructional features. This work addresses a current need in the early education field to develop a new generation of measurement identifying the active ingredients of prekindergarten classrooms that best support children's development (Burchinal, 2017; Weiland, 2018; Weiland & Guerrero-Rosada, 2022). Further, we inform a practical need of exploring whether time use is a potential lever for quality measurement and improvement. Our findings provide evidence that time use descriptions and profiles in prekindergarten classrooms appear to be measure dependent. Below, we explain implications of our key findings across comparisons of time use descriptions, the instructional profiles derived from the measures, and associations with children's language, math, and executive function gains.

Discussion

Time Use Descriptions of Prekindergarten Classrooms

Comparing time use descriptions and profiles across studies in the current literature is challenging due to variation in measurement strategies. We contrasted two intensive coding protocols with differences in time sampling strategies and level of coding (whole classroom vs. individual children) using the same set of classroom videos. By doing so, we held constant curricula, children's characteristics, day of observation, observation length, and instructional quality. Our design provides evidence that the time use measure adopted drives differences in time use descriptions.

Consistent with prior research, classrooms in our study measured with either the NR or ISI spent approximately 40% of time in whole-group activities (Cabell et al., 2013; Early et al., 2010; Justice et al., 2021) and 20% of time in transitions (Nores et al., 2022). However, time spent in the remaining 40% of time differed by measure. This remaining 40% of time corresponds to time when children's participation in learning settings is more likely to vary at the individual level, namely centers, small groups, and individual learning setting. As explained in our literature review, findings from prior studies are inconsistent regarding the proportion of time children spend in these settings. Our study shows that the lack of a clear pattern may at least in part be attributable to measurement choices that differ across studies.

This finding has two important implications. First, prior work using the Emerging Academic Snapshot has shown that descriptive results (i.e., means) may differ when measures are aggregated or analyzed at the child level instead of the classroom level (Chien et al., 2010; Early et al., 2010). We provide evidence that measurement protocols that rely on sampling children (i.e., ISI) and protocols that rely on sampling time (i.e., NR) are only consistent when used to describe time use occurring in whole group settings. A new direction for the field is to better understand the time and child sampling conditions that lead to consistent descriptions across instruments. Second, activities conducted during small groups and free choice/center settings respond to different instructional purposes. For example, tier-two interventions meant to address the specific needs of struggling and advanced students via intensive individual instruction often occur in small group arrangements (Dickman, 2006). A consensus on how to better represent instructional support during individual, centers, and small group learning settings is necessary to accurately capture the quality of prekindergarten experiences.

Our findings are consistent with studies showing that prekindergarten classrooms devote most time to language and literacy instruction (Bratsch-Hines et al., 2019; Burchinal et al., 2021; Justice et al., 2021; Nores et al., 2022; Weiland et al., 2023). Studies measuring time at the child-level using time-sampling strategies (i.e., snapshots) have found that children spend between 14% and 19% of their time in language/literacy instruction. Data collected and analyzed at the *child-level* for this study, using a second-to-second approach, showed that children spent 35% of time learning language/literacy content (Weiland et al., 2023). Our results —focused on the classroom-level— show similar patterns. We found that children spent 20% of time in language/literacy according to the NR (collected and analyzed at the classroom level) and 28% of time in these domains according to ISI (collected at the child-level and analyzed at the classroom level). Prior research has shown large variability in time spent in math (range = 4%–20%), science (range = 2%–11%), and social studies (1%–20%), a pattern we do not find in our study (see Table 1.1 for details). The instruments we compare measure these content areas consistently.

Time Use Profiles

Regarding our second research question—comparing time use profiles based on the NR and the ISI—we found that both instruments effectively differentiate *Whole Group Academic* (*WGA*) classrooms, an instructional profile extensively described in prior literature (Chien et al., 2010; Justice et al., 2021). However, only 48% of classrooms are assigned to this profile by both instruments. Both instruments also effectively identified *Balanced Mixed* (*BM*) classrooms in ways consistent with prior literature (Chien et al., 2010; Fuligni et al., 2012; Justice et al., 2021). However, we find when data are collected at the classroom level with the NR, the proportion of time children spend in language and literacy instruction is underestimated. Descriptions based on ISI data for the equivalent profile suggest this underestimation by the NR occurs because, in this profile, language and literacy are taught in combination with other content areas (i.e., in mixed or integrated content). This measurement choice has potential implications for predictive research, given evidence that language/literacy instruction tends to be more effective when contextualized in content-rich activities in prekindergarten (Maier et al., 2022).

Consistent with prior research (Cabell et al., 2013; Nores et al., 2022; Pianta et al., 2018), we show that time use profiles in our sample are linked to process and structural features of classroom quality. The direction of associations, however, is contradictory with the prior literature. Cabell and colleagues (2013) found the highest Instructional Support scores among classrooms spending more time in science activities, and lower scores among classrooms spending more time in language and math. Nores and colleagues (2022) found the opposite pattern with respect to classrooms' Emotional Support. Our results suggest that the direction of the association with quality depends on the instrument. For example, based on NR profiles, *WGA* classrooms have lower scores in Emotional Support and higher scores in Instructional support than *BM* classrooms; and based on ISI profiles, *WGA* classrooms have lower Emotional Support and higher Emotional Support scores than *BM* classrooms. As we discussed in the prior section, measurement error in estimated time at the variable level ultimately results in unreliable profiles regarding learning formats, and only partially reliably profiles regarding content areas.

Finally, we also examined whether profiles differ by children's baseline skills and classroom demographic composition simultaneously—considering teachers may offer different learning opportunities based on children's baseline skills (Sameroff, 2009; Weiland et al., 2023) or based on their own implicit and explicit beliefs and biases (Alvidrez et al., 1999; Robinson-Cimpian, Lubienski, Ganley, & Copur-Gencturk, 2014). To our knowledge, this is the first study showing there is no evidence of differences in children's skills by classroom instructional profiles, but there are differences in classrooms' demographic composition. In other words, teachers' choices about learning settings and content do not respond to children's skills at the beginning of the prekindergarten year but are related to the proportion of children eligible for free or reduced priced lunch, dual language learners, and proportion of Black and Latino children in the classroom, for both measures. A better understanding of these differences and the role these differences play in predictive models is an important new direction for the field.

Predictive Properties of time use Profiles

Finally, the NR and ISI resultant profiles did not differentially predict gains in children's academic and cognitive skills, with just one model predicting differential math gains—likely a spurious result given the number of models we fit. Because our resultant profiles did not replicate other work in prekindergarten classrooms (Chien et al., 2010; Fuligni et al., 2012) we are unable to contrast these predictive results with prior research. Despite this limitation, our models show evidence that children's differential gains between profiles decrease and are no longer statistically significant when we controlled for classroom quality, and subsequently, for classroom demographic composition.

Our study has several important limitations. First, our small sample size limits the generalizability of the profiles we estimated. We are unable to rule out whether the small group of classrooms that formed a third profile for each instrument (three *SGC / Mixed* when profiles were obtained using NR data and five *Whole Group / Mixed* when profiles were obtained using ISI data) are statistically significantly different from other classrooms. We were also unable to include children attending these classrooms in our predictive models. Research with a larger sample of classrooms is necessary to explore whether these profiles are consistently different from *WGA* and *BM* profiles and identify differential gains among them.

Second, we do not compare whether individual versus aggregated ISI measures lead to consistent results. We opted to conduct an apples-to-apples comparison by holding constant the unit of analyses so that we could make inferences related to the unit of the data collection. It is also possible that ISI results vary depending on the proportion of children observed for each classroom, a direction that new measurement work could explore in order to optimize measures that require a child-sampling strategy.

Third, our research is restricted to one school district, limiting our external validity. Large scale measurement initiatives with pre-registered agreements regarding children and time sampling strategies, observation protocol, and analytical approach are necessary to disentangle the association between these and other features of classroom instruction that better predict children’s developmental gains.

Despite these limitations, our results suggest that time use, a seemingly relatively transparent feature of classrooms, is measure dependent, related to quality, and related to classroom composition. Considering the relevance of adequately monitoring dosage for curriculum implementation and offering equitable opportunities to engage with rich and varied content, a stronger measurement consensus in the field is needed to be able to accurately determine instructional profiles that better predict children’s developmental gains during their prekindergarten year.

References

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

Bassok, D., Dee, T., & Latham, S. (2017). *The Effects of Accountability Incentives in Early Childhood Education* (No. w23859; NBER Working Paper Series). National Bureau of Economic Research. https://doi.org/10.3386/w23859

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. https://doi.org/10.1016/j.ecresq.2018.10.001

Brunsek, A., Perlman, M., Falenchuk, O., McMullen, E., Fletcher, B., & Shah, P. S. (2017). The relationship between the Early Childhood Environment Rating Scale and its revised form and child outcomes: A systematic review and meta-analysis. *PLOS ONE*, *12*(6), 1–29. https://doi.org/10.1371/journal.pone.0178512

Burchinal, M. (2017). Measuring early care and education quality. *Child Development Perspectives*, *12*(1), 3–9. https://doi.org/10.1111/cdep.12260

Burchinal, M., Garber, K., Foster, T., Bratsch-Hines, M., Franco, X., & Peisner-Feinberg, E. (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. https://doi.org/10.1016/j.ecresq.2020.10.005

Cabell, S. Q., DeCoster, J., LoCasale-Crouch, J., Hamre, B. K., & Pianta, R. C. (2013). Variation in the effectiveness of instructional interactions across preschool classroom settings and learning activities. *Early Childhood Research Quarterly*, *28*(4), 820–830. https://doi.org/10.1016/j.ecresq.2013.07.007

Chien, N. C., Howes, C., Burchinal, M., Pianta, R. C., Ritchie, S., Bryant, D. M., Clifford, R. M., Early, D. M., & Barbarin, O. A. (2010). Children's classroom engagement and school readiness gains in prekindergarten: Classroom engagement and school readiness gains. *Child Development*, *81*(5), 1534–1549. https://doi.org/10.1111/j.1467-8624.2010.01490.x

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.

Clements, D. H., & Sarama, J. (2011). Early childhood mathematics intervention. *Science*, 333(6045), 968-970.

Connor, C. M., Morrison, F. J., Fishman, B. J., Ponitz, C. C., Glasney, S., Underwood, P. S., Piasta, S. B., Crowe, E. C., & Schatschneider, C. (2009). The ISI Classroom Observation System: Examining the literacy instruction provided to individual students. *Educational Researcher*, *38*(2), 85–99. https://doi.org/10.3102/0013189X09332373

Dickman, G. E. (2006). RTI and reading: Response to intervention in a nutshell. In *Perspectives on language and literacy, Special Conference Edition*.

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., Barbarin, O. A., Winn, D.-M. C., Crawford, G. M., Frome, P. M., Clifford, R. M., Burchinal, M., Howes, C., Bryant, D. M., & Pianta, R. C. (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*(2), 177–193. https://doi.org/10.1016/j.ecresq.2009.10.003

Farran, D., Meador, D., Keene, A., Bilbrey, C., & Vorhaus, E. (2015). *Advanced Narrative Record Manual*. Vanderblit University, Peabody Research Institute.

Farran, D. C., & Bilbrey, C. (2014). *Variation in Observed Program Characteristics across Classrooms in the Tennessee Voluntary Pre-Kindergarten Program.* Society for Research on Educational Effectiveness.

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. https://doi.org/10.1016/j.ecresq.2011.10.001

Gehlbach, H., & Robinson, C. D. (2018). Mitigating illusory results through preregistration in education. *Journal of Research on Educational Effectiveness*, *11*(2), 296–315. https://doi.org/10.1080/19345747.2017.1387950

Gormley Jr, W. T., 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., Sachs, 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. https://doi.org/10.1016/j.ecresq.2020.07.009

Justice, L. M., Jiang, H., Purtell, K. M., Lin, T.-J., & Ansari, A. (2021). Academics of the early primary grades: Investigating the alignment of instructional practices from pre-K to third grade. *Early Education and Development*, *33*(7), 1–19. https://doi.org/10.1080/10409289.2021.1946762

Lerkkanen, M.-K., Kiuru, N., Pakarinen, E., Poikkeus, A.-M., Rasku-Puttonen, H., Siekkinen, M., & Nurmi, J.-E. (2016). Child-centered versus teacher-directed teaching practices: Associations with the development of academic skills in the first grade at school. *Early Childhood Research Quarterly*, *36*, 145–156. https://doi.org/10.1016/j.ecresq.2015.12.023

Maier, M. F., Hsueh, J., & McCormick, M. (2020). Rethinking Classroom Quality: What We Know and What We Are Learning. *MDRC*.

Maier, M. F., McCormick, M. P., Xia, S., Hsueh, J., Weiland, C., Morales, A., Boni, M., Tonachel, M., Sachs, J., & Snow, C. (2022). Content-rich instruction and cognitive demand in prek: Using systematic observations to predict child gains. *Early Childhood Research Quarterly*, *60*, 96–109. https://doi.org/10.1016/j.ecresq.2021.12.010

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., Burchinal, M., Early, D. M., & Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development*, *79*(3), 732–749. https://doi.org/10.1111/j.1467-8624.2008.01154.x

McCormick, M. P., Pralica, M., Guerrero-Rosada, P., Weiland, C., Hsueh, J., Condliffe, B., Sachs, J., & Snow, C. (2021). Can center-based care reduce summer slowdown prior to kindergarten? Exploring variation by family income, race/ethnicity, and dual language learner status. *American Educational Research Journal*, *58*(2), 420–455. https://doi.org/10.3102/0002831220944908

McCormick, M. P., Weiland, C., Hsueh, J., Maier, M., Hagos, R., Snow, C., Leacock, N., & 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. https://doi.org/10.1016/j.ecresq.2019.06.012

Morris, P. A., Mattera, S. K., & Maier, M. F. (2016). Making Pre-K Count: Improving math instruction in New York City. *MDRC*.

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. https://doi.org/10.1016/j.ecresq.2021.09.011

Perlman, M., Falenchuk, O., Fletcher, B., McMullen, E., Beyene, J., & Shah, P. S. (2016). A systematic review and meta-analysis of a measure of staff/child interaction quality (the Classroom Assessment Scoring System) in early childhood education and care settings and child outcomes. *PLOS ONE, 11*(12), e0167660. https://doi.org/10.1371/journal.pone.0167660

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System: Manual K-3*. Baltimore: Paul H Brookes.

Pianta, R. C., Mashburn, A. J., Downer, J. T., Hamre, B. K., & Justice, L. (2008). Effects of web-mediated professional development resources on teacher–child interactions in pre-kindergarten classrooms. *Early Childhood Research Quarterly, 23*(4), 431–451. https://doi.org/10.1016/j.ecresq.2008.02.001

Pianta, R. C., Whittaker, J. E., Vitiello, V., Ansari, A., & Ruzek, E. (2018). Classroom process and practices in public Pre-K programs: Describing and predicting educational opportunities in the early learning sector. *Early Education and Development*, *29*(6), 797–813. https://doi.org/10.1080/10409289.2018.1483158

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

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,* 1262–1281.

Rosenberg, J., Beymer, P., Anderson, D., van Lissa, C. j., & Schmidt, J. (2018). tidyLPA: An R Package to Easily Carry Out Latent Profile Analysis (LPA) Using Open-Source or Commercial Software. *Journal of Open Source Software*, 3(30), 978. https://doi.org/10.21105/joss.00978

Rosenthal, E. N., Riccio, C. A., Gsanger, K. M., & Jarratt, K. P. (2006). Digit Span components as predictors of attention problems and executive functioning in children. *Archives of clinical neuropsychology*, 21(2), 131-139.

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

Spurk, D., Hirschi, A., Wang, M., Valero, D., & Kauffeld, S. (2020). Latent profile analysis: A review and "how to" guide of its application within vocational behavior research. *Journal of Vocational Behavior*, *120*, 103445. https://doi.org/10.1016/j.jvb.2020.103445

van Cann, R., Jansen, S., Brinkkemper, S., van Cann, R., Jansen, S., & Brinkkemper, S. (2013). Noldus Information Technology. Software Business Start-up Memories: Key Decisions in Success Stories, 57-65.

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

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. *MDRC*.

Weiland, C., Moffett, L., Rosada, P. G., Weissman, A., Zhang, K., Maier, M., Snow, C., McCormick, M., Hsueh, J., & Sachs, J. (2023). Learning experiences vary across young children in the same classroom: Evidence from the individualizing student instruction measure in the Boston Public Schools. *Early Childhood Research Quarterly, 63*, 313–326. https://doi.org/10.1016/j.ecresq.2022.11.008

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

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: The Journal of the Association for Public Policy Analysis and Management*, *27*(1), 122-154.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). Woodcock-Johnson III tests of achievement.

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

Zamarro, G., Engberg, J., Saavedra, J. E., & Steele, J. (2015). Disentangling disadvantage: Can we distinguish good teaching from classroom composition? *Journal of Research on Educational Effectiveness*, *8*(1), 84–111. https://doi.org/10.1080/19345747.2014.972601

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

Table 1.1 Percent of Time Spent in Learning Settings and Activities

Measure	**Emerging Academic Snapshot**				**BCS**	**LISn**		**C-SNAP**	**EduSnap**	**ISI**
Study	Chien et al., 2010	Early et al., 2010	Fuligni et al., 2012	Cabell et al., 2013	Pianta et al., 2018	Bratsch-Hines et al., 2019	Burchinal et al., 2021	Justice et al., 2021	Nores et al., 2022	Weiland et al., 2023
Protocol and sample										
Range of observation time	NR	NR	NR	2.–4 h	1–4 h	3 h	2 h	NR	6 h	2–6 h
N (classrooms /children)	701/2,966	652/2061	125/206	346/NA	126/1506	63/366	63/366	*46/285	264/NR	37/263
N observations/class	2	2	2	1	2–3	1	1	4	2	2
Geography	Multistate (SWEEPS Study)		Los Angeles County, CA	See note # 4	Fairfax County, VA	Rural North Carolina		Two large Midwest districts	Philadelphia & New Jersey	Boston
Child or classroom level	Child	Child	Class	Class	Class	Class	Class	Child	Class	Child
Targeted parts of day	Whole day; non-specific		Morning, non-specific	Morning; non-specific	Morning; non-specific	Morning; language and instructional activities		NR; non-specific	Morning; non-specific	Lang/ literacy, math
Learning settings [1]										
Teacher-assigned (WG/SG/Ind)	--	37.00	--	--	--	--	--	--	--	--
Whole and large group	27.00	--	17.00	36.90	28.00	35.00	36.50	41.80	25.00	43.23
Small group	6.00	--	11.00	3.50	6.00	8.00	8.40	27.60	6.00	5.71
Dyad	--	--	--	--	--	--	--	14.80	--	--
Individual	4.00	--	--	0.80	4.00	--	--	14.90	3.00	16.13
Meals/Routines	--	34.00	--	--	--	--	--	--	--	27.50
Meals	12.00	--	13.00	13.10	13.00	--	--	--	9.00	--
Routines	21.00	--	16.00	12.30	19.00	--	--	--	--	--
Transitions	--	--	--	--	--	--	--	--	18.00	--
Free choice/Center	30.00	29.00	40.00	31.90	30.00	49.00	47.30	49.40[2]	40.00	35.21
Other/Recess/Outdoors	--	--	24.00	1.20	--	--	0.80	0.80	--	--
Learning activities [3]										
Academics	--	--	--	--	35.00	--	--	--	--	--
Language/Literacy	19.00	17.00	--	31.00	--	46.00	28.30	14.40	30.00	35.05
Math	8.00	8.00	--	6.80	--	--	16.30	3.80	19.00	20.46
Science	11.00	11.00	--	6.10	--	--	--	3.40	9.00	2.13
Social Studies	15.00	15.00	--	18.60	--	--	--	1.40	20.00	1.08
Socioemotional	--	--	--	--	4.00	--	--	--	--	0.55
Art, music, dance	15.00	15.00	--	21.80	14.00	--	--	1.90	15.00	9.50
Fine motor	10.00	10.00	--	--	--	--	--	--	--	--
Gross motor	6.00	6.00	--	--	--	--	--	--	13.00	0.63
Management	--	--	--	--	--	--	--	12.20	--	--
No content/Other	--	44.00	--	10.30	--	--	--	10.80	--	--

Other content	--	--	--	--	25.00	--	--	--	--	14.40
No content	--	--	--	--	22.00	--	--	--	--	--

Note.

[1] Learning settings are mutually exclusive and add up to 100%, except in Fuligni et al., 2012 – where outside time was double coded with another learning setting; Bratsch-Hines et al, 2019 and Burchinal et al., 2021 do not report time spent in individual learning; and Weiland et al., 2023– where "meals and routines" is accounted as a learning activity instead of a setting.

[2] "Free choice/center" is conceptualized as an activity in Justice et al., 2022. All other learning settings add up to ~100%.

[3] Learning activities only add up to 100% in the BCS (Pianta et al., 2018), since activities are double coded in other instruments.

BCS = Behavioral Coding System, LISn = Language Interaction Snapshot. EAS = Emerging Academics Snapshot (Ritchie, Howes, Kraft-Sayre, & Weister, 2001); C-SNAP = Classroom Snapshot (Ritchie et al., 2001).

Time sampling strategies: EAS: A minimum of 30 and a maximum of 50 1-minute snapshots per child during a program morning. BCS: 4 cycles across the morning, each included 10 intervals of 1 minute. LISn: 10 intervals of 30 seconds per child. C-SNAP: two separate 20-min cycles, each of which consisted of 20 one-minute intervals. Edu-Snap: four-minute cycles during which each child was observed and then coded for 1 min.

[4]Geography for Cabell et al. study: New York, NY; Hartford, CT; Chicago, IL; Stockton, CA; Dayton and Columbus, OH; Memphis; TN; Charlotte, NC; Providence, RI.

[5] Authors report that ¼ of the total sample of children were observed with BCS.

Table 1.2 Time Use Distribution in Prekindergarten by Classrooms' Instructional Profile

Study and Measure	**Chien et al., 2010-EAS**				**Fuligni et al., 2012-EAS**		**Justice et al., 2021–C-SNAP**			
Method	Latent Class Analysis (at the child level)				Latent Class Analysis		Latent Profile Analysis			
Evidence of variation by demographic characteristics	Yes, by race/ethnicity, socio-economic status, and household size.				Not explored		Yes. Differences by grade level and quality			
Profiles	Free play	Individual Instruction	Group Instruction	Scaffolded learning	Free choice	Structured Balanced	Individual L&L	**Whole Class L&L**	Whole Class Discussion	**Academic-light group work**
% Classrooms	51%	9%	27%	13%	29%	71%	11%	**50%**	6%	**33%**
Learning settings										
Whole group	20%	27%	36%	34%	10%	19%	19%	**53%**	75%	**21%**
Small group	4%	4%	11%	6%	4%	13%	16%	**21%**	17%	**58%**
Individual	2%	21%	3%	3%	--	--	62%	**23%**	6%	**19%**
Meals	14%	11%	10%	12%	--	--	--	**--**	--	**--**
Routines	20%	25%	26%	16%	--	--	--	**--**	--	**--**
Free Choice	41%	13%	15%	29%	61%	32%	--	**--**	--	**--**
Outdoor time	--	--	--	--	36%	--	--	**--**	--	**--**
Learning activities										
Language and Literacy	15%	27%	20%	34%	--	--	67%	**45%**	30%	**22%**
Math	6%	10%	10%	11%	--	--	7%	**12%**	39%	**11%**
Science	11%	9%	1%	16%	--	--	--	**--**	--	**--**
Social studies	17%	8%	11%	21%	--	--	--	**--**	--	**--**
Art	14%	17%	15%	16%	--	--	--	**--**	--	**--**
Motor development	17%	20%	13%	16%	--	--	--	**--**	--	**--**

Note. Learning settings are mutually exclusive and add up to 100% in Chien et al., 2010; 97%–98% in Justice et al., 2021; and are not mutually exclusive in Fuligni et al., 2012. Profiles in Justice et al., 2021 are estimated including classrooms from prekindergarten to third year. However, prekindergarten classrooms only were part of bolded profiles, 28% were classified in the Whole Class—Language and Literacy profile, and the remaining 72% were classified in the Academic Light Group Work profile.

Table 1.3 Children and Classroom Characteristics

	Mean or %	SD	% Missing
Child-level characteristics			
Demographic characteristics			
Female	51.82	--	0
Eligible for Free or Reduced Lunch (FRPL)	58.70	--	0
Dual Language Learner (DLL)	54.65	--	0
Asian Pacific Islander	19.83	--	0
Black	21.45	--	0
Latino	26.72	--	0
Other	6.07	--	0
White	25.91	--	0
Fall measures			
PPVT – Raw Score	73.05	28.29	3.64
PPVT – Standard Score	97.31	24.28	3.64
WJ Applied Problems Raw Score	12.64	5.01	3.64
WJ Applied Problems Standard Score	105.00	14.83	3.64
Forward Digit Span	3.15	1.03	3.64
Spring measures			
PPVT – Raw Score	87.55	26.95	6.47
PPVT – Standard Score	102.65	19.63	6.47
WJ Applied Problems Raw Score	15.88	4.47	6.88
WJ Applied Problems Standard Score	106.90	13.42	6.88
REMA Raw Score	17.28	8.73	6.47
REMA T Score	-2.51	1.14	6.47
REMA IRT	37.53	5.66	6.47
Forward Digit Span	3.52	1.05	6.47
Parents characteristics			
High-school degree	34.34	--	5.67
Two-years degree	20.17	--	5.67
Bachelor’s degree	15.45	--	5.67
Graduate degree	30.34	--	5.67
Adult has full-time job	88.98	--	4.45
Married	63.56	--	4.45
Mothers’ age at first birth	27.43	6.64	5.67
Respondent parent’s age	36.18	7.56	4.68
Household size	4.29	1.25	5.26
Classroom-level characteristics			
Structural and process quality			
Has a masters’ degree	80.00	--	0
Years of PreK experience	9.60	7.56	
CLASS–Instructional Support Score	3.16	0.60	0
CLASS–Classroom Organization	5.37	0.57	0
CLASS–Emotional Support	5.50	0.59	0
Classroom demographic composition			
Female	47.91	11.75	0
Eligible for Free or Reduced Lunch (FRPL)	62.66	24.45	0
Dual Language Learner (DLL)	51.60	25.55	0
Asian Pacific Islander	17.08	21.76	0

Black	25.46	21.92	0
Latino	31.81	20.35	0
Other	4.95	5.93	0
White	20.70	21.04	0

Note. *N* for child-level characteristics =247, *N* for classroom-level characteristics = 35.

Table 1.4 Differences Between Narrative Record and Individualizing Student Instruction

	NR (*SD*)	**ISI (*SD*)**	**Diff (*p* value)**
Learning settings[1]			
Whole group	0.41 (.13)	0.43 (.14)	-0.02 (*p* = 0.361)
Small group	0.02 (.05)	0.06 (.07)	-0.04 (*p* = 0.018)
Individual	--	0.15 (.10)	--
Centers/Small group	0.12 (.15)	--	--
Centers	0.22 (.15)[1]	0.36 (.12)	-0.14 (*p* = 0.000)
Transitions/Management	0.22 (.09)	0.20 (.05)	0.03 (*p* = 0.060)
Learning activities[2]			
Language/Literacy	0.20 (.09)	0.28 (.08)	-0.08 (*p* = 0.000)
Math	0.13 (.10)	0.12 (.07)	0.00 (*p* = 0.706)
Other content areas	0.09 (.07)	0.11 (.07)	-0.02 (*p* = 0.102)
Art, music, dance	0.03 (.03)	0.06 (.04)	-0.03 (*p* = 0.001)
Science	0.02 (.04)	0.02 (.04)	-0.00 (*p* = 0.688)
Social Studies	0.01 (.03)	0.02 (.05)	-0.00 (*p* = 0.918)
Socioemotional	0.01 (.02)	0.00 (.01)	0.01 (*p* = 0.092)
Motor development	0.01 (.01)	0.00 (.01)	0.00 (*p* = 0.106)
No content / Unknown content[3]	0.22 (.09)	0.15 (.10)	0.07 (*p* = 0.012)
Integrated or Mixed content	0.36 (.14)	0.42 (.14)	-0.06 (*p* = 0.106)

Notes.

[1] Learning settings add up to 100% of the time in the NR. In ISI, settings add up to 100% if "Transitions/Management" is excluded. We present Transitions as a learning setting to facilitate comparisons across instruments, but transitions are coded as a learning activity in ISI.

[2] Learning activities do not add up to 100%, since these are non-exclusive.

[3] "No Content" is used when no activity with instructional purpose is occurring per the NR protocol. "Other or unknown content" is used to indicate activities that do not have a clear instructional purpose and therefore cannot be categorized under the listed content areas per the ISI protocol.

Table 1.5 Fit Indices for Each Instrument's Profiles Solutions

	Log Likelihood	**Entropy**	**AIC**	**BIC**	**SABIC**	**BLRT (*p*)**	**Sizes**
Narrative Record							
1 profile	275.486	1	-514.973	-486.977	-543.181		
2 profiles	304.141	0.997	-552.282	-508.732	-596.162	0.01	3, 32
3 profiles	335.518	0.949	-595.036	-535.933	-654.587	0.01	3, 19, 13
4 profiles	349.453	0.981	-602.905	-528.249	-678.128	0.04	3, 16, 13, 3
ISI							
1 profile	330.126	1	-624.252	-596.255	-652.46		
2 profiles	340.499	0.785	-624.998	-581.448	-668.877	0.24	20, 15
3 profiles	352.661	0.888	-629.323	-570.219	-688.874	0.09	16, 14, 5
4 profiles	370.965	0.928	-645.93	-571.273	-721.152	0.01	16, 13, 5, 1

Note. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; SABIC = Sample Adjusted Bayesian Information Criterion; BLRT (*p*) = *p* value for the Bootstrapped Likelihood Ratio Test (BLRT).

Table 1.6 Distribution of Time in Learning Settings and Content Areas by Profile Membership (Fraction of Time)

	Narrative Record						ISI					
	SGC / Mixed (*N* = 3; 8.6%)		**Whole Group/ High Academic** (*N* = 19, 54.3%)		**Balanced/Mixed** (*N* = 13, 37.1%)		**Balanced/ Moderate Content** (*N* = 16; 45.7%)		**Whole Group/ High Academic** (*N* = 14, 40.0%)		**High Whole Group/Mixed** (*N* = 5, 14.3%)	
	Mean	***SD***	***Mean***	***SD***	***Mean***	***SD***	***Mean***	***SD***	***Mean***	***SD***	***Mean***	***SD***
Whole Group	0.27	0.05	0.52	0.07	0.29	0.07	0.34	0.12	0.49	0.10	0.59	0.07
Centers	0.05	0.06	0.18	0.13	0.32	0.13	0.40	0.13	0.31	0.11	0.33	0.04
Small Group	0.00	0.01	0.04	0.04	0.00	0.01	0.06	0.08	0.08	0.07	0.03	0.04
SG/Centers	0.54	0.04	0.07	0.07	0.09	0.09	--	--	--	--	--	--
Individual	--	--	--	--	--	--	0.21	0.10	0.13	0.08	0.05	0.06
Transitions	0.14	0.02	0.19	0.05	0.30	0.10	0.22	0.04	0.16	0.02	0.22	0.03
Language	0.15	0.06	0.25	0.09	0.14	0.06	0.24	0.06	0.34	0.06	0.27	0.05
Math	0.06	0.04	0.15	0.10	0.11	0.09	0.12	0.09	0.13	0.06	0.11	0.04
Other content	0.03	0.06	0.11	0.06	0.05	0.05	0.11	0.05	0.08	0.05	0.19	0.07
Mixed content	0.62	0.10	0.30	0.10	0.39	0.13	0.41	0.13	0.37	0.10	0.64	0.10

Table 1.7 Differences in Quality, Child and Family Demographics, and Classroom Characteristics by Profile Membership

	(1) **NR Balanced Mixed**	**(2)** **ISI Balanced Mixed**
Indicators of Classroom Process and Structural Quality		
CLASS–Emotional Support	0.48**	-0.40***
	(0.14)	(0.11)
CLASS–Classroom Organization	0.04	0.24
	(0.17)	(0.12)
CLASS–Instructional Support	-0.48***	0.45***
	(0.08)	(0.07)
Proportion of teachers with a Masters' degree	-0.04	0.33***
	(0.09)	(0.07)
Teacher years of PreK experience	-0.00	0.03***
	(0.01)	(0.00)
Average class size at school	-0.04*	-0.05***
	(0.02)	(0.01)
Children's Baseline Skills and Age		
PPVT score in fall of 2016	0.00	-0.00
	(0.00)	(0.00)
WJ-AP score in fall of 2016	0.00	-0.02**
	(0.01)	(0.01)
DS score in fall of 2016	-0.06	0.04
	(0.04)	(0.03)
Age in fall of 2016	-0.03	0.03
	(0.11)	(0.09)
Parents		
Mother's age	0.01	0.00
	(0.01)	(0.00)
Size of household	-0.02	0.03
	(0.03)	(0.02)
Proportion of adults who work	-0.01	0.17*
	(0.11)	(0.07)
Proportion of married adults	-0.16*	-0.12*
	(0.08)	(0.06)
Proportion with a high-school degree	0.03	-0.01
	(0.10)	(0.07)
Proportion with BA+	0.05	0.08
	(0.10)	(0.07)
Proportion with income less than 25000	-0.04	-0.08
	(0.09)	(0.06)
Classroom composition		
Proportion of children eligible for free or reduced lunch	0.92***	-2.13***
	(0.25)	(0.22)
Proportion of girls	-0.70	-0.33
	(0.41)	(0.34)
Proportion of dual language learners	0.69*	1.78***
	(0.33)	(0.24)
Proportion of Latino children	-2.28***	-0.91**

	(0.42)	(0.31)
Proportion of Black children	0.24	1.87***
	(0.29)	(0.23)
Proportion of Asian children	-1.26***	-0.34
	(0.33)	(0.26)
Proportion of mixed or other race children	0.32	2.43***
	(0.60)	(0.44)
Constant	0.22	0.53
	(0.91)	(0.66)
Observations	166	153
R-squared	0.48	0.76

Note. ****p* < 0.001; ***p* < 0.01; **p* < 0.05. Whole–Group Academic profile is the reference group for both instruments. *N* for the NR = 221 (excluding 26 children in the Small Group / Centers Mixed Profile and *N* for ISI = 216 (excluding 31 children in the High Whole Group / Mixed Profile).

Table 1.8 Differences in Children's Gains by Profile Membership

	PPVT			**WJ-AP**			**REMA**			**Digit Span**		
NR Models												
Balanced/Mixed	3.37	2.38	1.68	-0.97*	-1.06*	-0.67	-0.38	-0.75	0.35	-0.17	-0.14	-0.16
	(2.48)	(2.70)	(3.01)	(0.42)	(0.46)	(0.50)	(0.79)	(0.87)	(0.95)	(0.14)	(0.15)	(0.17)
ISI Models												
Balanced/Mixed	-2.45	-3.54	-1.02	0.65	0.60	0.80	-1.86*	-1.41	-1.46	-0.05	-0.01	0.17
	(2.31)	(2.65)	(3.88)	(0.49)	(0.53)	(0.71)	(0.84)	(0.95)	(1.38)	(0.14)	(0.16)	(0.23)
Child and family demographics	X	X	X	X	X	X	X	X	X	X	X	X
Process and structural quality covariates		X	X		X	X		X	X		X	X
Classroom composition covariates			X			X			X			X

Note. The reference group is the: *Whole Group / High Academic* for both measures. Standard Errors in parenthesis. ****p* < 0.001; ***p* < 0.01; **p* < 0.05. Models include random intercepts for classrooms and schools.

Chapter 2 - Study 2: Teacher-Reported Complexity of Instruction Does Not Predict Children's Within-Grade Math and Language Gains in Prekindergarten, Kindergarten, or First Grade

Often known as "fade-out" or "convergence," a well-known phenomenon in the early childhood education field is that children's prekindergarten gains tend to diminish as they move through elementary school (Abenavoli, 2019; Bailey et al., 2017). The reasons for this phenomenon are unknown, with mixed evidence from empirical studies (Unterman & Weiland, 2020). Recently, researchers have hypothesized that *alignment* of children's learning experiences is a key lever for sustaining children's gains. The concept of alignment encompasses both horizontal alignment of standards, curricula, and assessments within settings and vertical alignment of experiences across the prekindergarten to the third-grade period (Harding et al., 2020; Stipek et al., 2017). Under this hypothesis, a possible cause for fade-out is that children are repeatedly exposed to content they have already mastered, or activities in kindergarten and early elementary years are too easy for their skills repertoire (Bassok, Latham, et al., 2016; Claessens et al., 2014; Engel et al., 2013; Justice et al., 2021; Vitiello et al., 2020).

Examining children's exposure to advanced content may be an avenue for improving schools' horizontal alignment (e.g., between practices and standards) and vertical alignment (e.g., standards coverage across years). Instructional alignment, which we approach as the extent to which children are consistently exposed to developmentally appropriate instructional content in the early grades is a school and system-level malleable factor relevant for policy and practice. . Research using nationally representative data has shown that, during kindergarten, basic content in math, language, and literacy is taught more frequently than advanced content, often teaching children what they already know (Bassok et al., 2016; Claessens et al., 2014; Engel et al., 2013). These findings are consistent with those from observational studies examining instructional alignment in large school districts, which have shown that, generally, teachers focus on basic content across prekindergarten and the early elementary school grades (Justice et al., 2021; Vitiello et al., 2020). These examples of instructional misalignment may relate to the fact that most children in the U.S. are not proficient in math and reading in the fourth grade, with large disparities by family income and race/ethnicity (Hussar et al., 2020).

However, determining what is "basic" and "advanced" presents measurement challenges, with little clear guidance in the literature around alternative approaches or definitions to date. To identify the proportion of basic versus advanced content that children experience, researchers largely have aimed to quantify different indicators of instructional practices as more advanced and basic. For example, Claessens and colleagues (2014) leveraged the strict alignment between children's academic skills as measured by nationally representative assessments (National Center for Educational Statistics–NCES) and learning standards to a) identify the skills that most children had mastered by the beginning of the kindergarten year, b) categorize activities as "basic" or "advanced" depending on the proportion of children who had mastered each skill by the beginning of kindergarten (as approached by Engel et al., 2013). Using this approach, based on children's proficiency levels in ECLS-K and ECLS-K:2011 surveys, researchers found consistent evidence that frequent exposure to basic content is associated with smaller math gains (Claessens et al., 2014; Engel et al., 2013, 2016) and exposure to advanced content is associated with larger reading and language gains (Claessens et al., 2014; Crosnoe et al., 2016). However, these definitions may create generalizability challenges due to differences in the curriculum implemented by varied programs. For example, a content item such as “Matching small sets (up to 5 objects) with the corresponding numerals” might be deemed basic if the equivalent skill is tested in a program where most children have had enough instruction to perform correctly by prekindergarten entry, but the same item could be deemed advanced if tested in another sample. In other words, "advanced" is endogenous to the proficiency of children tested; making challenging for practitioners and policymakers to use this approach to improve alignment.

In this study, we examine whether children’s within-year language and math gains are predicted by two alternative approaches to defining the complexity of language/literacy and math instruction. In the first approach, we define complexity of content conceptually based on learning standards and calculate classrooms’ average-grade level of instruction. Learning standards, as defined by national and state-level policies, indicate the scope and sequence of content to which children are exposed; classrooms’ *average grade level of instruction* reflects teachers’ adherence to said sequence. In the second approach, we define complexity of content empirically based on the level of difficulty of each content as estimated by one-parameter IRT models (also known as Rasch models). In other words, we obtained parameters of difficulty based on the probability of each item being taught across the prekindergarten, kindergarten, and first grade years by teachers in the sample and describe these distributions by year.

Our study makes several contributions to the literature. First, we conducted our study in partnership with the Boston Public Schools, a national leader in attempting to align children’s early learning experiences across prekindergarten to second grade through a set of curriculum and professional development reforms (McCormick et al., 2020). Second, we look beyond a single grade and consider children’s within-grade gains across three years, in prekindergarten, kindergarten, and first grade —a period that considers the critical transition to elementary school.

This broader perspective helps identify how the average grade level of instruction varies over these three critical years, all of which are of interest to practitioners looking for strategies to support children's early learning. Finally, to our knowledge, our conceptual and empirical approaches for defining complexity of instruction at the grade level is novel in this literature and accordingly makes a measurement contribution to this emerging area of research.

Descriptive Evidence on Exposure to Basic and Advanced Content and Redundancy Across Years

The notion of advanced content builds on the concept of vertical alignment, or the extent to which learning standards, curriculum, and assessment from each grade level serve as a foundation on which to build the standards, curriculum, and assessment of the following grade (Franko et al., 2018; Harding et al., 2020; McCormick et al., 2020; Stipek et al., 2017). From this perspective, there is an explicit sequence in which foundational knowledge is taught before more complex knowledge is introduced. This approach is consistent with evidence showing that pre-defined developmental progressions are necessary for building a foundation for mastering more complex academic skills (Foorman et al., 2016; Frye et al., 2013). This rationale has been successfully used in curriculum development based on learning trajectories where children master some skills first before accessing additional levels of challenge through subsequent instruction (Clements et al., 2013).

Some authors have used a conceptual approach to examine such alignment. In a sample of 117 public prekindergarten classrooms and 295 public kindergarten in Virginia, Vitiello and team (2020) collected teacher reports of literacy and math content taught as part of general classroom instruction throughout the school year. Literacy and math content experts identified whether each item was most appropriate for prekindergarten, kindergarten, first, or second grade;

and then the authors cross-walked experts’ responses with the state standards local to the school district and consulted again with the experts to resolve any discrepancies. By doing so, they identified one set of items as most appropriate for prekindergarten (basic) and other items as most appropriate for kindergarten (advanced). Their findings show that both prekindergarten and kindergarten teachers taught most of the basic items. In literacy, prekindergarten teachers reported they taught 90% of basic items and kindergarten teachers reported they taught 93% of basic items. A similar pattern was observed in math (91% prekindergarten and 96% kindergarten). In other words, kindergarten teachers reported teaching significantly more basic items than prekindergarten teachers (3.167%, p = 0.002 for literacy and 2.99%, p = 0.003 for math). For example, basic literacy content, such as understanding conventions of print and matching letters to sounds; and math content, such as subitizing, matching small sets and correspondence between numbers and quantity, were equally likely to be taught by teachers at both grade levels. Regarding the set of advanced items, approximately half were taught in prekindergarten (48% for literacy and 65% of math) and the majority were taught in kindergarten (95% for literacy and 93% for math). Some examples of advanced literacy items are blending separate sounds of a word; identifying words, sentences, ending punctuation; and advanced math content such as solving addition problems and subtracting single-digit numbers were taught significantly more by kindergarten teachers than prekindergarten teachers. In conclusion, prekindergarten teachers reported they taught the material expected for the grade level, and almost all the material expected for kindergarten. Kindergarten teachers also reported they taught the material expected for the grade level and repeated approximately half of the prekindergarten material. These results suggest there was a substantial overlap between the literacy and math content taught at both grade levels.

Other authors have used an empirical approach to determine whether instructional content is basic or advanced. For example, in cross-sectional work, Engel and colleagues (2013) used secondary nationally representative data to identify activities that matched closely with students' math proficiency levels at kindergarten entry. Then, they described how frequently these activities were taught (measured in days per month of the school year). Engel and colleagues (2013) showed that teachers spent considerably more time on Basic Counting and Shapes, content that corresponds with math Proficiency Level 1, which 95% of sample students had already mastered when they entered kindergarten, than they did on any of the other math content measures (i.e., Patterns and Measurement, Place Value and Currency, and Addition and Subtraction). Teachers spent the least amount of time on the content more closely linked with the highest level of children's proficiency, namely Addition and Subtraction.

When exposure to advanced or basic content is examined across grade-levels —as Vitiello and colleagues did in their conceptual approach— empirically-approached studies can also highlight misalignment across grade levels, in which children receive redundant instruction across years. Evidence from the North Carolina Prekindergarten Program showed that not only there is significant repetition or redundancy in basic math and language/literacy content from prekindergarten to kindergarten, but also this redundancy is more likely for children from families who live at or below the poverty line (Cohen-Vogel et al., 2021). Cohen-Vogel and colleagues (2021) used teachers reports of activities that were their major focus of instruction for the school year to create a measure of redundancy by identifying the amount of overlap between activities done in both prekindergarten and kindergarten. Using data from 63 PreK and 145 kindergarten classrooms in rural North Carolina, the authors found that 37% of language/literacy content (39% in reading, 42% in writing, and 46% in language) and 32% of math content (18%

operations, 24% measurement, 33% geometry, and 62% numeracy) was redundant across years. Children from families at or below the poverty line were more likely to experience redundancy, especially in language and literacy (b = 0.028, $p < 0.05$).

Researchers have posited that repetition of content from PreK to Kindergarten might play a role in prekindergarten fade-out or catch up (Jenkins et al., 2018; M. McCormick et al., 2017), in which the skills of children who attended PreK and those that did not converge partially or fully as soon as the end of kindergarten. However, some repetition may be necessary to support children's development; there are no agreed-upon or research-based thresholds of repetition that support versus inhibit early learning.

Associations Between Advanced Instructional Content and Children's Academic Gains

Studies have also shown that frequent exposure to advanced content is associated with children's academic gains in kindergarten, illustrating its potential importance in children's learning (Claessens et al., 2014; Crosnoe et al., 2016; Engel et al., 2013; Jenkins et al., 2018; Justice et al., 2021). In literacy, teacher reports of more days per month teaching advanced reading (i.e., skills that the majority of children had not mastered by kindergarten entry) during kindergarten were associated with larger gains in the ECLS-K reading assessment ($b = .053$, $p < .01$; Claessens et al., 2014). More recently, in the context of a curriculum intervention, researchers found a negative association ($ES = .09$) between basic reading activities and children's outcomes at the end of kindergarten, which appears to be concentrated in children's receptive vocabulary (PPVT). In contrast, during first grade, there was a positive association between advanced language and literacy activities and children's gains in early writing skills (Jenkins et al., 2018).

In math, teacher reports of more days per month teaching advanced math were positively associated with larger gains in ECLS-K math assessment scores (*b* = .065, *p* < .01). In contrast, reports of additional days per month working on basic math content were associated with smaller gains in the same math assessment (*b* = –.041, *p* < .01; Claessens et al., 2014). Consistently, analyses by content area showed that each additional day per month learning basic counting and shapes was negatively associated with math gains (-0.02 *SD*), whereas each additional five days per month learning place value and currency content was associated with larger math gains (0.03 *SD*); and each additional four days per month learning addition and subtraction content was associated with larger math gains (0.04 *SD*) (Engel et al., 2013, 2016). In sum, evidence suggests that exposure to advanced content —or fewer exposure to basic content— across grade levels is positively associated with children's academic gains, perhaps because classrooms consistently introducing advanced content decrease non-essential repetition. However, we have yet to learn whether systematic exposure to advanced content benefits all children, and whether positive associations with gains are robust to the varied ways how exposure to advanced content is operationalized.

Present Study

In this study, we extend the existing literature by examining association between the complexity of classroom instruction and children's within-grade gains in a diverse sample of prekindergarten, kindergarten, and first grade students in the Boston Public Schools. To do so, we use two alternative measures of complexity, one conceptually defined based on national and state standards that are not sample-dependent and one empirically defined based on a parametric estimation of difficulty obtained from our sample survey respondents. We add to prior conceptually oriented work by using national and state standards as a policy reference, which can aid replication and cross-sample comparisons. We add to prior empirically oriented work by accounting for the level of difficulty of each item across grade levels, which accounts for items that are meant to be taught across the spectrum of early grades. Finally, we extend these analyses to include first grade. We aim to answer the following research questions:

1. What is the average complexity of language/literacy and math instruction in prekindergarten, kindergarten, and first grade? How does it differ using conceptually based versus empirically based measures?
2. Do conceptually and empirically based measures of content complexity predict children's within-grade prekindergarten, kindergarten, and first grade language and math gains?

Method

Participants and Setting

The sample consists of 579 children across the 2016–2017, 2017–2018, and 2018–2019 school years. For the first year of the study, the sample included 401 prekindergartners nested in 51 classrooms, 20 BPS schools, and 11 community-based organization (CBO) partner centers. For the second year, we recruited 178 additional students and 71 children left the study, for a total sample of 508 kindergartners nested in 102 classrooms in 54 schools. For the third year, 86 children left the study, which resulted in a sample of 422 first graders nested in 51 classrooms in 23 schools. On average, 68% of students in the sample were eligible for free or reduced-price lunch, 53% of students were Dual Language Learners (DLL), 28% were Black, 21% were White, 32% were Hispanic, 16% were Asian, and 3% were mixed race or another race. About 40% of third-grade students in study schools met or exceeded expectations on the 2015–2016 state English/Language Arts exam, and 45% met or exceeded expectations on the state math exam.

Although the schools in the sample are generally representative of the population of BPS elementary schools offering a prekindergarten program, our sample had lower proportions of Black students (32% at the district level) and higher proportions of students meeting or exceeding expectations on the 2015–2016 ELA exam (36% at the district level). On average, teachers in the sample were experienced teaching at their grade level (PreK Mean =14.79 *SD* = 8.77, K Mean =12.62 *SD* = 8.48, 1st grade Mean = 13.76, *SD* = 8.90), and most of them had a master's degree (PreK = 71%, K = 82%, 1st grade = 87%). In Boston Public Schools, teachers are required to hold a master’s degree. However, we include this control to account for potential differences in education with prekindergarten classrooms in CBOs, where teachers were exempt from this requirement at the time of this study.

Procedures

The Institutional Review Boards at the lead organization for this study approved the human subjects plan prior to the commencement of study activities.

School and Classroom Recruitment

The research team randomly selected 25 schools from the full set of 76 schools that offered the public prekindergarten program and 11 CBOs in Boston implementing the BPS prekindergarten model. Four schools declined participation and one was designated as a pilot school for developing new measures. All prekindergarten teachers assigned to general education or inclusion classrooms in each of the 20 participating sample schools and 11 CBOs were invited to participate in the study in the fall of 2016. Ninety-six percent of teachers in 19 schools (*N* = 41 classrooms) and 11 CBOs (*N* = 11 classrooms) agreed to the study activities. The research team followed sample children into public kindergarten and first grade, and, accordingly, asked their teachers in those grades to participate in the study. 95% of kindergarten teachers gave consent to participate in the study activities.

Teacher Survey

In the spring of 2017, 2018, and 2019, respectively, we asked prekindergarten, kindergarten, and first grade teachers to complete a survey reporting on the content they taught as a major focus of instruction during the school year, their demographic characteristics, and their teaching experience. Teachers were given a list of instructional practices selected and adapted from the ECLS-K Spring Classroom Instruction Questionnaire (Tourangeau et al., 2015) and the Common Core Standards (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010). Teachers were asked to check whether each practice was part of their general classroom instruction, or if it was taught only to selected students. For practices that were part of teachers' general instruction, they checked whether each was a major or minor focus of instruction. We structured the survey questions this way to limit social desirability bias—i.e., a teacher could indicate they did teach a practice to a minority of students (which was not a focus of our study). Out of 205 teachers in the study sample, 82.44% responded to the survey across years (PreK = 92%, K = 77%, and 1st grade = 82%).

Classroom Videotaped Observations

During the Winter of 2017, 2018, and 2019, each classroom was videotaped for two hours during two visits scheduled in advance with teachers. The research team 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. Classrooms' process quality was assessed in video recordings using the

Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2008). Prior to coding videos with the CLASS (Pianta et al., 2008), coders participated in a two-day training and then established reliability on a set of master codes created by the CLASS developers. As recommended by the measure's protocol (Pianta et al., 2008), coders used cycles of 20 minutes for observing and 10 minutes for scoring, which they repeated up to four times for each observation. Coding began when instruction commenced in the video and ceased after 80 minutes of observed time. We double-coded 20% of the observations to assess interrater reliability. Throughout the coding process, we conducted drift checks wherein observers had to code a master tape every three weeks to ensure they stayed reliable across time. 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.

Student Recruitment and Direct Assessments

The research team began recruitment activities in late September 2016 and completed them by late November 2016. Additional children enrolled in kindergarten and first grade were recruited in September 2017 and 2018, respectively. From late September 2016 through late November 2016, the research team solicited informed consent for all students in participating classrooms via backpack mail, providing an overview of the study and a blank consent for the parent to complete and return to the child's classroom. Overall, 81% of children in participating classrooms agreed to participate. The research team randomly selected 50% (~6–10 per classroom) of consented children to participate in student-level data collection activities. We trained research staff to achieve reliability and then collected direct assessments of academic skills in the fall of 2016 and 2017 for baseline measures for prekindergarten and kindergarten, respectively, and spring of 2017, 2018, and 2019 for outcome measures. We used the Pre-language Assessment Scale (preLAS; Duncan & DeAvila, 1998) Simon Says and Art Show tests to determine the administration language for a subset of assessments (Barrueco et al., 2012). Of the 377 children assessed in prekindergarten, 43 (11%) completed a subset of assessments in Spanish in fall 2016. Of the 483 children assessed in kindergarten, 15 (3%) completed assessments in Spanish in fall of 2017. All children were assessed in English during first grade. Details on the assessment missingness, sample composition and size are presented in Table 2.1.

Parent Surveys

We used text messages and emails to contact the consenting parents of all students who were selected for the study sample and collected parental demographic information via 20-minute surveys in the fall of prekindergarten and kindergarten, and the spring of first grade. The research team sent biweekly reminders to complete the survey, and paper copies were sent via backpack mail to collect outstanding surveys. In addition to English (96% respondents), surveys were available in Spanish (6% respondents), Mandarin (2% respondents), and Vietnamese (1% respondents). All parents received a $25 gift card for completing the survey. Parents completed the survey for 88% of the 579 children in the current study sample.

Measures

Average Grade Level and Complexity of Classroom Instruction

We used a set of survey items drawn from the ECLS-K Spring Classroom Teacher Questionnaire (NCES, 2011) and the Common Core Standards (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) to measure our key question predictors, complexity of classroom instruction. In all, there were 57 items in Pre-K (language/literacy = 31, math = 26), 64 in kindergarten (language/literacy = 32, math = 32), and 100 in first grade (language/literacy = 54, math = 46). For each item, teachers indicated whether they taught specific language, literacy, and math practices, whether each practice was a major focus of instruction during the year, and whether it was taught to all or selected students. We used the indicators of practices that teachers reported were major focus of instruction for the year only.

From teachers' responses to these items, we first constructed a conceptually based measure. To do so, we coded each practice to indicate the expected grade-level of the activities in the Common Core Standards for English Language Arts and Math (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) and the Massachusetts State Standards (Massachusetts Department of Elementary and Secondary Education, 2017), which includes prekindergarten standards. For each study year, items corresponding to the current grade level were centered at zero, items one year below and above the grade level were scored with -1 and 1 respectively, and we followed the same rule to code practices that were up to two years below or above the grade level. Then, we averaged across items to obtain an average grade level of instruction based on the practices each teacher reported to have covered as part of their major focus for the year, where a score of zero represents a classroom where the teacher focused on practices that correspond to their current grade level (on average), positive scores correspond to teachers who focused on practices above their grade level (on average), and negative scores correspond to teachers who focused on practices below their grade level (on average). We did so within three language domains (i.e., Language and Reading Comprehension, Literacy Foundational Skills, and Oral and Written Composition) and two math domains (Numeracy, and Content Specific Math which includes Operations, Geometry, and Measurement). See Table 3C in Appendix C for alpha reliability and details on the number of items confirming each scale, and Figures 2C and 3C for distributions of each measure across years.

Next, we created an *empirically based measure* in relation to the practices taught by other teachers in the sample across grades. To do so, we used 1-parameter Item Response Theory (IRT) models to estimate a θ score representing the complexity of instruction. IRT models are a measurement strategy designed to quantify at which point of a continuum latent trait (θ, representing an unobservable characteristic or attribute) a given individual performs (Briggs, 2008; Kim & Camilli, 2014; Linden, 2016). We used this strategy to identify teachers' position in the latent trait of interest—complexity of instruction—based on the items they reported to teach. A score of zero corresponds to a teacher whose practices are situated at the midpoint of complexity for their school year, negative scores correspond to teachers including easier practices in comparison with other teachers in their same school year, and positive scores correspond to teachers including more difficult practices in comparison with other teachers in the same school year. We started by estimating Rasch models including the same items per subscale as we did in our conceptual approach, and then a) removed uninformative items based on their probability functions and Item Characteristic Curves (ICCs), b) tested the items fit for each 1-parameter model and removed non-fitting items, and c) tested each model for unidimensionality and estimated the corresponding theta scores for classrooms in the sample. See Appendix D, Table 4D, for final scale-level ICCs and distributions for each and across years in Figures 4D and 5D.

As a robustness check, we also created dichotomized versions of each of our measures by assigning a value of 1 to classrooms where the complexity of instruction was above the mean in relation to their grade level and a value of 0 to classrooms where the complexity of instruction was below the mean in relation to their grade level.

Children's Language and Math Skills

We assessed children's receptive language skills using the Peabody Picture Vocabulary IV (PPVT IV; Dunn & Dunn, 2007). The PPVT IV is a nationally normed measure that has been used widely in diverse samples of young children, has excellent split-half and test–retest reliability estimates, and strong validity properties (Dunn & Dunn, 2007). It requires children to choose which of four pictures best represents a stimulus word, verbally or non-verbally. We used the raw score total as our outcome measure in our primary analysis. The research team assessed all children on the PPVT—regardless of whether they passed the PreLAS language screener—to describe an equivalent measure of receptive language skills in English across the full sample.

We assessed children's math skills using the Woodcock-Johnson III Applied Problems subtest (WJ-AP III; Woodcock et al., 2001, 2005) and the Research-based Early Mathematics Assessment (REMA; Clements & Sarama, 2008). The Woodcock–Johnson Applied Problems subtest requires children to perform relatively simple calculations and solve arithmetic problems. Its estimated test–retest reliability for 2- to7-year-old children is 0.90 (Woodcock et al., 2001). It has been nationally normed and used with diverse populations of children (Gormley, Gayer, Phillips, & Dawson, 2005; Wong, Cook, Barnett, & Jung, 2008). The research team assessed Spanish-speaking children who did not pass the PreLAS language screener using Batería III Woodcock Muñoz (Woodcock, Munoz-Sandoval, Ruef, & Alvarado, 2005), which follows similar norms to the Woodcock–Johnson English version and allows for combining scores across both English and Spanish in the sample. We used raw scores for our main models.

We also measured math skills using the REMA (Clements & Sarama, 2011), an assessment of a broader range of children's early math skills (e.g., numeracy, geometry, operations, spatial reasoning). The alpha reliability of the test subscales ranges from r = 0.71 (geometry) to 0.89 (numeracy). We present results using the REMA raw score in our main analyses. We did not assess children on the REMA during the fall of prekindergarten, and thus used the Woodcock–Johnson Applied Problems score as a baseline for the prekindergarten model examining REMA gains.

Classroom Process Quality and Indicators of Structural Quality

We included measures of process and structural quality as covariates. We coded general 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. Emotional Support is a composite measure of four subscales—positive and negative climate, sensitivity and regard for student's perspective. Classroom Organization includes measures of behavior management, productivity, and instructional learning formats. Instructional Support includes concept development, language modeling, and quality of feedback. All the dimensions are directly scored on a 7-point scale, except for negative climate which is reverse-coded. In prior work with our prekindergarten sample, the CLASS did not predict gains in children's language, math, and working memory (Guerrero-Rosada et al., 2021).

Measures of teachers' experience and education were constructed based on survey data. Teachers reported their highest level of education, from which we created an indicator of whether the teacher holds a masters' degree. They also reported the years of experience teaching in prekindergarten, which we used as a continuous measure.

Parent Characteristics

We constructed indicators of the reporting parents' level of educational attainment as a proxy for socio-economic status (high-school, two-years degree, bachelor's degree, graduate degree); whether there was at least one parent in the home working full-time (35 hours/week or more); and whether the parent was married or lived with a partner. We also used continuous measures to describe the age of the child's mother at her first birth, the parent respondent's age in the fall of 2016, and the number of people living in the household. We include these to match prior work with this sample (Guerrero-Rosada, Weiland, McCormick, et al., 2021; M. P. McCormick et al., 2020, 2021; Weiland et al., 2023); experts have advised including the same covariates across studies from the same dataset to prevent illusionary results (Gehlbach & Robinson, 2017).

Children's Demographic Characteristics

We accessed administrative records from the school district on children's demographic characteristics to create indicators of their gender; eligibility for free or reduced lunch; dual language learner status (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); race/ethnicity (Asian or Asian American, Black, Latinx/Hispanic, other or mixed race, and White) and birthdate, which we used to calculate children's age when their baseline measures were collected.

Analytical Approach

To identify *the average complexity of language/literacy and math instruction in prekindergarten, kindergarten, and first grade and how does it differ using conceptually based and empirically based measures* (RQ 1), we estimated means for each grade and then used t-tests to compare prekindergarten versus kindergarten and kindergarten vs first grade, for each measure.

To examine *whether conceptually and empirically based measures of content complexity predict children's within-grade prekindergarten, kindergarten, and first grade language and math gains* (RQ2), we estimated regression models separately within grade using random intercepts for schools, as shown in Equation 1:

$$Outcome_{ijk} = \beta_0 + \beta_1 Instruction_{jk} + \beta_2 Baseline_{ijk} + \chi_{ijk} + \rho_{ijk} + \lambda_{jk} + \mu_k + \varepsilon_{ik} \quad (1)$$

where the subscript *i* refers to an individual student, *j* denotes an individual classroom, and *k* represents an individual school. $Outcome_{ijk}$ refers to children's spring scores in vocabulary and numeracy, $Instruction_{jk}$ is the key question predictor of average grade level of instruction or complexity of instruction, and $Baseline_{ijk}$ is the child's corresponding baseline score on the outcome. Because we did not collect data on children's skills in the fall of first grade, we used scores during the spring of kindergarten as baseline measures to estimate their corresponding gains over a full year rather than an academic year (which is time period of focus for prekindergarten and kindergarten). χ_{ijk} is a vector of student-level characteristics including their race/ethnicity, gender, DLL status, eligibility for free or reduced lunch, age, testing interval from Fall to Spring, and an indicator of whether the child attended PreK at BPS or a CBO. $\rho_{ijk}$is a vector of parent characteristics measured at the child level including indicators for whether the parent works, is married, completed a two-year degree, completed a bachelors' degree, or completed a graduate degree (with parents who completed high school as the reference group), responding parent's age, mother's age at first birth, and the number of people in the home. λ_{jk} is a set of characteristics measured at the classroom level that includes the CLASS Instructional Support score, teacher experience, and an indicator of whether the teacher has a masters' degree, which we included to disentangle profiles based on the grade level or complexity of instruction from process and structural quality. Models include random intercepts for schools, μ_k and ε_{ijk} are the school and student-level residual terms.

Results

Descriptive Statistics

Our sample was racially, linguistically, and socio-economically diverse (see child-level characteristics in Table 2.1). CLASS scores, on average, ranged from low to moderate in Instructional Support and from moderate to high in Emotional Support and Classroom Organization. These scores varied across years (see Table 2.2). Instructional Support was moderate in prekindergarten, low during the kindergarten year (-0.69 *SD*, $p < 0.001$), and moderate again for first grade (0.47 *SD*, $p < 0.001$). Emotional Support and Classroom Organization were high throughout the three study years, but Classroom Organization scores were higher for the kindergarten year (0.42 *SD*, $p < 0.001$). Emotional Support scores were also higher for the kindergarten year (0.23 *SD*, $p < 0.05$), and lower for the first grade (-0.30 *SD*, $p < 0.01$).

RQ 1: What is the Average Complexity of Language/Literacy and Math Instruction in Prekindergarten, Kindergarten, and First Grade; and How Does it Differ Using Conceptually Based Versus Empirically Based Measures?

Table 2.2 shows our conceptually and empirically based scores for each focal grade. Based on conceptually based measures (centered at zero to represent a clear correspondence with learning standards), the average grade level of language and reading comprehension was 1.67 in prekindergarten (roughly one and a half year above grade level), 0.50 in kindergarten (half year above grade level) and 0.03 in 1st grade (at grade level). In other words, children received language and reading comprehension instruction with the same level of complexity across years when defined in relation to learning standards. The average grade level of instruction for Literacy Foundational Skills was almost a year above grade level in prekindergarten (0.69) and half year above grade level in kindergarten (0.46). In first grade, the average grade level of this domain decreased to -0.07—slightly below grade level. Only 11 teachers reported implementing Oral and Written Composition practices in prekindergarten, and the average grade level of their instruction was 1.47—a year and a half above grade level. Given that our conceptually based measures are simple averages across coded items, we cannot generalize this level of instruction to other prekindergarten classrooms. In kindergarten, the average grade level of Oral and Written Composition was 0.62 —above grade level— and 0.05 in first grade —at grade level.

We also show the average grade level of math instruction in Table 2.2. The average grade level of Numeracy instruction was 0.43 in prekindergarten, 0.07 in kindergarten, and -0.25 in first grade. In other words, teachers taught content almost half a year above grade level in prekindergarten, at grade level in kindergarten, and below grade level in first grade. For Operations, Geometry, and Measurement instruction, the average grade level in prekindergarten was 1.17, one year above grade level. In kindergarten, the average grade level of instruction decreased to 0.18 —at grade level— and decreased once again for first grade to -0.32— below grade level. A visual examination of the overlap between the average grade level of instruction across years provides further context for these magnitudes in our conceptually based language/literacy and math instruction (see Appendix C, Figures 2C and 3C).

Based on empirical measures —parameters of complexity for each scale ranging between -2 and 2 (see Table 2.2) the average grade level of Language and Reading Comprehension was -0.04 in prekindergarten, decreased to -0.10 in kindergarten, and then slightly increased to 0.20 in first grade. Mean comparisons show that on average, the complexity of Language and Comprehension instruction remained the same across years. For Literacy Foundational Skills, the average grade level of instruction was -1.06 in prekindergarten, 0.24 in kindergarten, and 0.38 in first grade. These scores show there was a substantial increase in the complexity of instruction between prekindergarten and kindergarten (Mean increase = 1.30, $p < 0.001$), but not between kindergarten and first grade (Mean increase = 0.13, $p > 0.05$). In Language and Reading Composition (including all prekindergarten classrooms in a parametric approach), the average complexity of instruction was -1.03 in prekindergarten, 0.12 in kindergarten, and 0.49 in first grade. Both increases were statistically significant (mean increase = 1.15, $p < 0.001$; and 0.37, $p < 0.01$, respectively).

For math, the average grade level of numeracy instruction was statistically similar between prekindergarten and kindergarten (-0.34 and -0.16, respectively), and increased for the first grade (0.72, $p < 0.001$). In Operations, Geometry, and Measurement, the average grade level increased in 0.33 ($p < 0.05$) from prekindergarten to kindergarten, and 0.39 ($p < 0.05$) from kindergarten to first grade.

RQ 2: Do Conceptually and Empirically Based Measures of Content Complexity Predict Children's Within-Grade Prekindergarten, Kindergarten, and First Grade Language and Math Gains?

Multilevel models showed no associations between the average grade level of language/literacy and math instruction and children's language and math gains (see Table 2.3). Although associations did not reach conventional levels of statistical significance, we observed variation in their direction and magnitudes across grade levels in language. For example, while the coefficients associated to the grade level of Language and Reading Comprehension and

Literacy Foundational Skills were positive in prekindergarten, these coefficients are negative and slightly larger for the kindergarten year and are close to zero in 1st grade. In math (see Table 2.4), the direction of associations changes depending on the outcome measure and varies across school years. For example, prekindergarten models predicting Woodcock-Johnson gains show positive coefficients whereas the coefficient associated to the grade-level of numeracy is negative when predicting gains in REMA scores. We observed the opposite pattern for kindergarten. Then, for first grade, the coefficient associated to the grade level of both math domains was negative when predicting Woodcock-Johnson gains and positive when predicting gains in REMA scores.

Similarly, results from multilevel models using empirically based measures of content complexity also suggest null associations with language (see Table 2.3) and math (See Table 2.4) gains. Coefficients for language and literacy measures did not reach conventional levels of statistical significance and their magnitudes were close to zero, ranging between 0.01 and 0.02 standardized association in prekindergarten, between -0.01 and -0.04 standardized association in kindergarten, and between -0.00 and 0.02 standardized association in first grade (see Table 2.3). In math, coefficients were also small (near zero), with directions that varied across grade levels (see Table 2.4).

Robustness Checks

We estimated alternative models using dichotomized versions of our measures to examine whether results could be confounded by non-normal distributions in some measures and grade levels, as we show in Figures 2C and 3C for conceptually based measures in Appendix C, and 4D and 5D for empirically based measures in Appendix D. Robustness Checks are shown in Table 2.5 for language models and Table 2.6 for math models. Consistent with our main approach, results were null across models. In conceptually based measures, exceptions were a small association between Literacy Foundational Skills (*standardized association* = -0.05, *p* < 0.05) and Oral and Written Composition (*standardized association* = 0.09, *p* < 0.05) and children's vocabulary gains. In math, only one empirically based measure significantly and negatively predicted Woodcock Johnson gains in kindergarten (*standardized association* = -0.13, *p* < 0.05).

Discussion

In this paper, we aimed to describe the complexity of language/literacy and math instruction in prekindergarten, kindergarten, and first grade, and examine its association with within-grade children's language and math gains. We also examined the consistency of our results when using conceptually based versus empirically based measures. Our substantive contribution is motivated by the need to identify school-level malleable factors that play a role in ensuring that prekindergarten instruction is strong and subsequent experiences are likely to be sustain its contributions on children's development and academic achievement. Methodologically, we contribute two novel approaches to assess the complexity of instruction across grade levels, which we operationalized as grade-level of instruction. Our findings provide evidence that conceptually and empirically based measures of content complexity are discriminative across grade levels. However, these measures did not have predictive power in relation to children's language and math outcomes. Below, we explain implications of our key descriptive and predictive findings across measures.

Complexity of Instruction Across Grade Levels

Prior literature has examined exposure to "advanced" and "basic" content during kindergarten, motivated by the hypothesis that repetition of content is one of the mechanisms of skills fadeout (Claessens et al., 2014; Engel et al., 2016). More recently, this work has been extended to directly examine alignment and repetition between prekindergarten and kindergarten (Cohen-Vogel et al., 2021; Vitiello et al., 2020), and between kindergarten to first grade (Jenkins et al., 2018). We examined content complexity in prekindergarten, kindergarten, and first grade using two different measures. We discuss our findings in relation to the prior literature and our measurement approaches.

There are three main conclusions from our descriptive work. First, within each school year, classrooms varied in the complexity of content instruction—operationalized as the average grade level of instruction in relation to learning standards and as the degree of complexity as estimated by Rasch parameters. In relation to learning standards, our findings are consistent with similar work in Virginia and North Carolina (Cohen-Vogel et al., 2021; Vitiello et al., 2020), where prekindergarten teachers reported they focused on skills that are usually taught in kindergarten. Consistently with Vitiello and colleagues (2020) and Cohen-Vogel and colleagues (2021), teachers in Boston Public Schools Prekindergarten classes also reported teaching above grade level (1.18 overall language/literacy and 0.78 overall math) and kindergarten teachers taught at grade level on average (0.52 overall language/literacy and 0.12 overall math). The first contribution of our study is extending these descriptions to first-grade classrooms, where we observed a continued decrease in the grade level of instruction (0.00 overall language/literacy and -0.28 overall math). When examined in relation to learning standards, we found that the classrooms' average grade level of instruction decreases as the grade level increases, which suggests there is overlap in the instructional content across years. Our second measure, empirically based, allowed us to examine the change in content complexity across years in our study sample. Our findings show that complexity of instruction changed more from prekindergarten to kindergarten (*SD* = 0.95, p < 0.001) than it did from kindergarten to first grade (*SD* = 0.46, p < 0.05) in language and literacy, but displayed the opposite pattern in math. Specifically, prekindergarten and kindergarten classrooms had similar complexity of math instruction (*SD* = 0.30, *p* > 0.05) and complexity in math increased significantly at first grade (*SD* = 0.79, *p* < 0.001). Prior work in North Carolina has shown that, during kindergarten, 43% of language/literacy items and 32% of math items had already been taught in prekindergarten (Cohen-Vogel et al., 2021). A limitation of our work is that we did not directly assess repetition across grade levels, and therefore cannot rule out whether the similar levels of complexity we observed across some grade levels correspond to specific content that is being repeated as opposed to teachers covering different content that should have been taught in a different grade level.

Second, we also found that complexity of instruction varied among sub-domains (i.e., Language and Reading Comprehension, Literacy Foundational Skills, and Oral and Written Composition within Language/Literacy; or Numeracy and Operations, Geometry, and Measurement within Math), which suggests children are exposed to different content progressions. For example, our conceptually based measures showed that only 23% of classrooms included Oral and Written Composition content as a major focus of their instruction in prekindergarten, and teachers who covered this content domain reported practices that were expected for kindergarten and first grade. Additionally, there was almost a grade level of difference between the complexity of Language and Reading Comprehension (taught one year above grade level, on average) and Literacy Foundational Skills (taught at grade level, on average) in prekindergarten, but both domains were covered at approximately half-year above grade level in kindergarten, and exactly at grade level in first grade. A potential implication is that the complexity of instruction in Literacy Foundational Skills increases as children move through the early years, whereas the complexity of Language and Comprehension does not. In contrast, the average complexity of Numeracy, Operations, Geometry, and Measurement are almost identical within years, suggesting teachers are connecting instruction across content areas. These findings are relevant for schools assessing vertical alignment, and the field can move this area forward by developing accountability and observational instruments that can help systematically identify these differential content trajectories.

Third, we conclude that the average complexity of instruction consistently decreased from prekindergarten to kindergarten and from kindergarten to first grade, different to what is expected from a developmental perspective (Franko et al., 2018; Harding et al., 2020; McCormick et al., 2020; Stipek et al., 2017). Our conceptually based measures directly show the mentioned decrease in the average grade level of instruction for all assessed content domains—a consistent decrease from prekindergarten to first grade suggests that kindergarten and first grade experiences are not introducing additional challenges for children that could potentially help sustain their prekindergarten learning. Our empirically based measures, consistently, show the almost perfect overlap in the overall complexity of instruction for language and math. However, because the IRT measures maximize the information available for teachers who reported teaching only a few or no activities for a given area, we observe large increases in the complexity of the content domains that were barely taught in prekindergarten (i.e., Literacy Foundational Skills and Oral and Written Composition). More research is needed to understand the implications of these "content gaps" for kindergarteners, and strategies to better align the prekindergarten—kindergarten transition.

We used two measurement approaches to identify to what extent findings vary when the complexity of instruction is set in relation to the sample or in relation to a standard comparable across samples. This concept has been approached in prior literature as the frequency and the proportion of advanced versus basic content (Claessens et al., 2014; Engel et al., 2013, 2016; Vitiello et al., 2020). By identifying a classroom's average grade level of instruction in relation to learning standards, we show that prekindergarten classrooms introduce complex content, and such complexity systematically decreases in subsequent grade-levels. Although this approach has the advantage of allowing comparisons across samples, it also comes with limitations such as relying on a small subset of learning standards with different coverage across grades (see Appendix C for details on the measure composition). In contrast, using a parametric approach allowed us to identify a continuum of complexity based on the probability that teachers in the sample report covering each item, and estimate the specific position of each teacher through such continuum of estimated complexity. For some content domains, such as Language and Reading Comprehension and Numeracy, these measures consistently identify that the complexity of instruction remains constant from prekindergarten to kindergarten. For more specific content domains such as Literacy Foundational Skills, Operations, Geometry and Measurement, empirically based measures identify small increases in complexity from year to year, potentially reflecting the sequence and scope of the BPS curricula (McCormick et al., 2020). These differences suggest that measures of content complexity, or advanced versus basic content coverage, can reflect curriculum characteristics that are particular to studies' samples. Further research is needed to identify more replicable and generalizable approaches to examine the role of content complexity and its alignment across years on children's development. Both measures show that classrooms across the three examined grade levels implement practices with similar complexity (see Figures 2C and 3C in Appendix C, and Figures 4D and 5D in Appendix D).

Predictive Properties of Complexity of Instruction Measures

Finally, our conceptually and empirically based measures did not predict gains in children's language and math skills. Ours is the first study using the average grade-level and complexity of instruction to predict children's gains. Prior studies examining frequency of exposure to advanced content have observed associations between such frequency and children's reading gains (*b* = .053, *p* < .01; Claessens et al., 2014) and a negative association between the frequency of basic reading activities and children's vocabulary (*ES* = .09) in kindergarten (Jenkins et al., 2018). Jenkins and colleagues also found a positive association between advanced language and literacy activities and children's gains in early writing skills in first grade. We did not analyze academic (reading and writing) outcomes. However, we observed very small positive associations between complexity and children's prekindergarten (Effect Sizes = 0.04 and 0.03 in conceptually and empirically based measures, respectively) and first grade gains (0.02 in both measures), and a small negative association with children's gains in kindergarten (-0.04 and -0.03 in conceptually and empirically based measures, respectively). The sizes we detected were smaller than those detected in Jenkins et al (2018) for the same outcome—and did not reach conventional levels of statistical significance.

In math, prior research has posited that more days per month teaching advanced math were positively associated with larger gains in ECLS-K math assessment scores (*b* = .065, *p* < .01) and more days working on basic math content were associated with smaller gains in the same math assessment (*b* = –.041, *p* < .01; Claessens et al., 2014). A content domain that was deemed basic such as counting and shapes negatively predicted math gains (-0.02 *SD*), whereas content domains deemed advanced such as place value and currency, and addition and subtraction, were associated with larger math gains (0.03 *SD* and 0.04 *SD*) (Engel et al., 2013, 2016). Although the coefficients associated with our conceptually and empirically based measures did not reach conventional levels of statistical significance, we found a few interesting patterns. In prekindergarten, higher levels of complexity were associated with larger constrained and unconstrained math gains (sizes ranging between 0.02 and 0.04, except for numeracy which was 0.00). In kindergarten, higher levels of complexity were negatively associated with constrained math gains (sizes between -0.03 and -0.06, except for Operations, Geometry and Measurement which was 0.01). All associations with REMA gains were near zero. In first grade, we observed the same directions and magnitudes of effects, except for positive associations with REMA gains ranging between 0.02 and 0.10. Validating and examining these measures in different school districts and with larger sizes would help assess whether the magnitudes we observed have substantive meaning for the field.

Our study has several additional important limitations. First, we did not control for potential differences in content complexity associated with classroom composition. Prior work has shown that children's instructional time use varies across demographic groups and such variation has implications in predictive models (Guerrero et al., in prep). We expect to address this limitation by replicating this methodological approach, assessing potential differences in content complexity by classroom composition and including this additional set of covariates in predictive models (Zamarro et al., 2015). Second, we did not examine whether the average grade level of instruction responds to the baseline knowledge of children in the classroom. Although we do not have access to the baseline literacy and math skills for all children in the classrooms, we expect to address this limitation by estimating models using standardized outcomes for our language and math measures, replicating prior work conducted in the same prekindergarten and kindergarten sample (McCormick et al., 2021; Weiland et al., 2023). Third, we were unable to collect baseline measures in the fall of the first-grade year. We addressed this limitation in our modeling approach by using the spring of kindergarten measure as baseline for first grade models. This solution does not address the fact that children's skills can vary over the summer in relation to the experiences to which they have access (McCormick et al., 2021). A fourth limitation is that our small sample size limits the predictive power in our multilevel models. Fifth, we were also unable to examine literacy outcomes from prekindergarten to first grade. It is possible that outcomes closely aligned with instruction, as observed in prior studies could yield different results in relation to predictive models (Claessens et al., 2014; Crosnoe et al., 2016; Engel et al., 2013; Jenkins et al., 2018; Justice et al., 2021).

Despite these limitations, our results suggest that children from the same grade level are exposed to content with varied complexity, and such levels of complexity are not sustained through early years. Considering the relevance of vertical alignment for sustaining children's gains during elementary years, developing tools that schools can use to consistently and reliably assess content coverage and complexity throughout the years can help improve the quality of early education experiences.

References

Abenavoli, R. M. (2019). The mechanisms and moderators of “fade-out”: Towards understanding why the skills of early childhood program participants converge over time with the skills of other children. *Psychological bulletin, 145*(12), 1103.

Bailey, D., Duncan, G. J., Odgers, C. L., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of research on educational effectiveness, 10*(1), 7-39.

Barrueco, S. (2012). Assessing young bilingual children with special needs. In *Assessment of Young Children with Special Needs* (pp. 237-254). Routledge.

Bassok, D., Latham, S., & Rorem, A. (2016). Is kindergarten the new first grade? *AERA Open*, *2*(1), 233285841561635. https://doi.org/10.1177/2332858415616358

Briggs, D. C. (2008). Using explanatory item response models to analyze group differences in science achievement. *Applied Measurement in Education, 21*(2), 89–118. https://doi.org/10.1080/08957340801926086

Claessens, A., Engel, M., & Curran, F. C. (2014). Academic Content, Student Learning, and the Persistence of Preschool Effects. *American Educational Research Journal*, *51*(2), 403–434. https://doi.org/10.3102/0002831213513634

Clements, D. H., & Sarama, J. (2011). Early childhood mathematics intervention. *Science*, 333(6045), 968-970.

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.

Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E. (2013). Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies: Persistence of effects in the third year. *American Educational Research Journal, 50*(4), 812–850.

Cohen-Vogel, L., Little, M., Jang, W., Burchinal, M., & Bratsch-Hines, M. (2021). A missed opportunity? Instructional content redundancy in Pre-K and Kindergarten. *AERA Open, 7*, 233285842110061. https://doi.org/10.1177/23328584211006163

Crosnoe, R., Benner, A. D., & Davis-Kean, P. (2016). Preschool enrollment, classroom instruction, elementary school context, and the reading achievement of children from low-income families. In G. Kao & H. Park (Eds.), *Research in the Sociology of Education* (Vol. 19, pp. 19–47). Emerald Group Publishing Limited. https://doi.org/10.1108/S1479-353920150000019003

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

Dunn, L. M., & Dunn, D. M. (2007). PPVT-4: Peabody Picture Vocabulary Test. Pearson Assessments.

Engel, M., Claessens, A., & Finch, M. A. (2013). Teaching students what they already know? The (mis)alignment between mathematics instructional content and student knowledge in kindergarten. *Educational Evaluation and Policy Analysis, 35*(2), 157–178. https://doi.org/10.3102/0162373712461850

Engel, M., Claessens, A., Watts, T., & Farkas, G. (2016). Mathematics content coverage and student learning in kindergarten. *Educational Researcher, 45*(5), 293–300. https://doi.org/10.3102/0013189X16656841

Foorman, B., Beyler, N., Borradaile, K., Coyne, M., Denton, C. A., Dimino, J., Furgeson, J., Hayes, L., Henke, J., & Justice, L. (2016). Foundational skills to support reading for understanding in kindergarten through 3rd grade. Educator’s Practice Guide. NCEE 2016-4008. *What Works Clearinghouse*.

Franko, M. D., Zhang, D., & Hesbol, K. (2018). Alignment of learning experiences from prekindergarten to kindergarten: Exploring group classifications using cluster analysis. *Journal of Early Childhood Research*, *16*(3), 229–244. https://doi.org/10.1177/1476718X18775761

Frye, D., Baroody, A. J., Burchinal, M., Carver, S. M., Jordan, N. C., & McDowell, J. (2013). Teaching math to young children. educator’s practice guide. What Works Clearinghouse. NCEE 2014-4005. *What Works Clearinghouse*.

Gehlbach, H., & Robinson, C. D. (2017). Mitigating illusory results through pre-registration in education. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3025214

Gormley Jr, W. T., 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., Sachs, 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. https://doi.org/10.1016/j.ecresq.2020.07.009

Harding, J. F., McCoy, D. C., & McCormick, M. P. (2020). Understanding alignment in children’s early learning experiences: Policies and practices from across the United States. *Early Childhood Research Quarterly*, *52*, 1–4. https://doi.org/10.1016/j.ecresq.2019.12.007

Hussar, B., Zhang, J., Hein, S., Wang, K., Roberts, A., Cui, J., Smith, M., Mann, F. B., Barmer, A., & Dilig, R. (2020). The Condition of Education 2020. NCES 2020-144. *National Center for Education Statistics*.

Jenkins, J. M., Watts, T. W., Magnuson, K., Gershoff, E. T., Clements, D. H., Sarama, J., & Duncan, G. J. (2018). Do high-quality kindergarten and first-grade classrooms mitigate preschool fadeout? *Journal of Research on Educational Effectiveness*, *11*(3), 339–374. https://doi.org/10.1080/19345747.2018.1441347

Justice, L. M., Jiang, H., Purtell, K. M., Lin, T.-J., & Ansari, A. (2021). Academics of the early primary grades: Investigating the alignment of instructional practices from Pre-K to Third Grade. *Early Education and Development*, 1–19. https://doi.org/10.1080/10409289.2021.1946762

Kim, S., & Camilli, G. (2014). An item response theory approach to longitudinal analysis with application to summer setback in preschool language/literacy. *Large-Scale Assessments in Education*, *2*(1), 1–17. https://doi.org/10.1186/2196-0739-2-1

Linden, W. (2016). *Handbook of item response theory 1*. Chapman and Hall/CRC.

Maier, M. F., McCormick, M. P., Xia, S., Hsueh, J., Weiland, C., Morales, A., Boni, M., Tonachel, M., Sachs, J., & Snow, C. (2022). Content-rich instruction and cognitive demand in prek: Using systematic observations to predict child gains. *Early Childhood Research Quarterly*, *60*, 96–109. https://doi.org/10.1016/j.ecresq.2021.12.010

Massachusetts Department of Elementary and Secondary Education. (2017). *Massachusetts Curriculum Framework for English Language Arts and Literacy*. Authors.

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. *MDRC*.

McCormick, M. P., Pralica, M., Guerrero-Rosada, P., Weiland, C., Hsueh, J., Condliffe, B., Sachs, J., & Snow, C. (2021). Can center-based care reduce summer slowdown prior to kindergarten? Exploring variation by family income, race/ethnicity, and dual language learner status. *American Educational Research Journal*, *58*(2), 420–455. https://doi.org/10.3102/0002831220944908

McCormick, M. P., Weiland, C., Hsueh, J., Maier, M., Hagos, R., Snow, C., Leacock, N., & 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. https://doi.org/10.1016/j.ecresq.2019.06.012

National Governors Association Center for Best Practices, & Council of Chief State School Officers. (2010). *Common Core State Standards.*

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System: Manual K-3*. Paul H Brookes.

Snow, C. E., Burns, M. S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children*. National Academy Press.

Stipek, D., Franke, M., Clements, D., Farran, D., & Coburn, C. (2017). PK-3: What does it mean for instruction? *Social Policy Report*, *30*(2), 1–23. https://doi.org/10.1002/j.2379-3988.2017.tb00087.x

Tourangeau, K., Nord, C., Lê, T., Sorongon, A. G., Hagedorn, M. C., Daly, P., & Najarian, M. (2015). Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:

2011). User's Manual for the ECLS-K: 2011 Kindergarten Data File and Electronic Codebook, Public Version. NCES 2015-074. *National Center for Education Statistics*.

Vitiello, V. E., Pianta, R. C., Whittaker, J. E., & Ruzek, E. A. (2020). Alignment and misalignment of classroom experiences from Pre-K to kindergarten. *Early Childhood Research Quarterly*, *52*, 44–56. https://doi.org/10.1016/j.ecresq.2019.06.014

Weiland, C., Moffett, L., Rosada, P. G., Weissman, A., Zhang, K., Maier, M., Snow, C., McCormick, M., Hsueh, J., & Sachs, J. (2023). Learning experiences vary across young children in the same classroom: Evidence from the individualizing student instruction measure in the Boston Public Schools. *Early Childhood Research Quarterly*, *63*, 313–326. https://doi.org/10.1016/j.ecresq.2022.11.008

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.

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: The Journal of the Association for Public Policy Analysis and Management*, *27*(1), 122-154.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). Woodcock-Johnson III tests of achievement.

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

Zamarro, G., Engberg, J., Saavedra, J. E., & Steele, J. (2015). Disentangling disadvantage: Can we distinguish good teaching from classroom composition? *Journal of Research on Educational Effectiveness*, *8*(1), 84–111. https://doi.org/10.1080/19345747.2014.972601

Table 2.1 Child and Family Demographics

	Mean or %	***SD***	**% Missing**
Demographic characteristics			
Female	51%	--	0%
Eligible for Free or Reduced Lunch (FRPL)	68%	--	0%
Dual Language Learner (DLL)	53%	--	0%
Asian Pacific Islander	16%	--	0%
Black	28%	--	0%
Latino	32%	--	0%
Other	3%	--	0%
White	20%	--	0%
Attended BPS PreK	58%	--	0%
Prekindergarten sample			
Fall measures Y1			
PPVT – Raw Score	72.07	27.44	6%
PPVT – Standard Score	85.33	26.63	9%
WJ Applied Problems Raw Score	11.96	5.17	7%
WJ Applied Problems Standard Score	103.1	14.76	7%
Spring measures Y1			
PPVT – Raw Score	85.33	26.63	9%
PPVT – Standard Score	100.42	10.52	1%
WJ Applied Problems Raw Score	14.92	5.00	9%
WJ Applied Problems Standard Score	104.36	14.23	9%
REMA Raw Score	15.65	8.69	9%
Kindergarten sample			
Fall measures Y2			
PPVT – Raw Score	87.57	27.7	1%
PPVT – Standard Score	98.22	11.5	1%
WJ Applied Problems Raw Score	15.83	5.27	1%
WJ Applied Problems Standard Score	100.89	15.48	1%
REMA – Raw Score	11.32	5.62	6%
Spring measures Y2			
PPVT – Raw Score	101.28	26.81	9%
PPVT – Standard Score	101.67	10.25	5%
WJ Applied Problems Raw Score	19.28	4.65	9%
WJ Applied Problems Standard Score	102.5	14.75	9%
REMA Raw Score	16.11	7.99	9%
First grade sample			
Spring measures Y3			
PPVT – Raw Score	120.33	23.77	26%
PPVT – Standard Score	103.58	10.49	20%
WJ Applied Problems Raw Score	23.92	3.98	26%
WJ Applied Problems Standard Score	100.21	12.62	26%
REMA Raw Score	17.81	8.06	26%
Parents characteristics			
High-school degree	33%	--	12%
Two-years degree	30%	--	12%
Bachelor’s degree	16%	--	12%
Graduate degree	21%	--	12%

Adult has a full-time job	87%	--	12%
Married	55%	--	12%
Mothers' age at first birth	26.32	6.69	14%
Respondent parent's age	36	7.16	13%
Household size	4.3	1.65	12%

Note. *N* for child-level characteristics = 401 PreK, 508 K, 422 1st grade.

Table 2.2 Classroom Complexity of Instruction and Structural/Process Quality

	Prekindergarten		**Kindergarten**		**1st Grade**			
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Diff K–PK**	**Diff 1st Grade–K**
Complexity of instruction (conceptually based measure)								
Overall Language/Literacy Instruction	1.18	0.26	0.52	0.18	0.00	0.15	-0.66***	-0.52***
Language and Reading Comprehension	1.67	0.22	0.50	0.21	0.03	0.19	-1.17***	-0.47***
Literacy Foundational Skills	0.69	0.39	0.46	0.27	-0.07	0.23	-0.23***	-0.27***
Oral and Written Composition	1.47	0.45	0.62	0.30	0.05	0.24	-0.85***	-0.56***
Overall Math	0.79	0.30	0.12	0.32	-0.27	0.23	-0.68***	-0.40***
Numeracy	0.43	0.33	0.07	0.44	-0.25	0.26	-0.36***	-0.32***
Operations, Geometry, Measurement	1.17	0.46	0.18	0.28	-0.32	0.26	-1.00***	-0.49***
Complexity of instruction (empirically based measure)								
Overall Language/Literacy	-0.70	0.60	0.09	0.62	0.36	0.59	0.78***	0.27*
Language and Reading Comprehension	-0.04	0.91	-0.10	0.82	0.20	0.83	-0.06	0.31
Literacy Foundational Skills	-1.06	0.71	0.24	0.67	0.38	0.58	1.30***	0.13
Oral and Written Composition	-1.03	0.52	0.12	0.74	0.49	0.64	1.15***	0.37**
Overall Math	-0.35	0.58	-0.10	0.83	0.46	0.70	0.25	0.55***
Numeracy	-0.34	0.52	-0.16	0.96	0.56	0.55	0.17	0.72***
Operations, Geometry, Measurement	-0.36	0.84	-0.03	0.87	0.36	1.06	0.33*	0.39*
Structural and process quality								
Has a masters' degree	71%	--	82%	--	87%	--	10%	5%
Years of PreK experience	14.79	8.77	12.62	8.48	13.76	8.9	-2.17	1.14
CLASS – Instructional Support	3.19	0.63	2.5	0.58	2.97	0.54	-0.69***	0.47***
CLASS – Classroom Organization	5.36	0.58	5.79	0.66	5.59	0.69	0.42***	-0.20
CLASS – Emotional Support	5.51	0.59	5.74	0.51	5.44	0.63	0.23*	-0.30**

Note. *N* for instruction, teacher education, and experience Y1= 48, except for Oral and Written Composition given that only 11 teachers reported implementing these practices, Y2 = 79, and Y3 = 42. *N* for CLASS Y1 = 50, Y2 = 51, and Y3 = 45. Standard errors in parentheses. ****p* < 0.001, ***p* < 0.01, **p* < 0.05

Table 2.3 Multi-Level Models Predicting Language Gains (PPVT Raw Scores)

	PK	K	1st Grade
Standards-based measures			
L&R Comprehension	4.304	-10.464	-1.704
	(5.381)	(5.972)	(4.091)
Literacy Foundational Skills	3.464	-4.397	1.570
	(3.028)	(4.628)	(3.014)
Oral and Written Composition	-4.752	1.510	2.664
	(5.893)	(3.500)	(3.443)
Language/Literacy (overall)	3.779	-6.614	2.566
	(4.678)	(7.123)	(5.150)
IRT measures			
L&R Comprehension	0.447	-1.230	-0.016
	(1.348)	(1.185)	(0.946)
Literacy Foundational Skills	0.884	-0.169	0.190
	(1.370)	(1.115)	(0.897)
Oral and Written Composition	0.660	-0.909	0.932
	(1.575)	(1.194)	(0.897)
Language/Literacy (overall)	1.019	-1.128	0.500
	(1.743)	(1.413)	(1.025)

Note. Standard errors in parentheses. ****p* < 0.001, ***p* < 0.01, **p* < 0.05

Table 2.4 Multi-Level Models Predicting Math Gains

	PK		**K**		**1st Grade**	
	WJ-AP	**REMA**	**WJ-AP**	**REMA**	**WJ-AP**	**REMA**
Standards-based measures						
Numeracy	0.202	-0.328	-0.514	0.300	-0.524	3.057
	(0.808)	(1.145)	(0.369)	(0.730)	(0.831)	(1.828)
Geometry, Algebra, Measurement	0.413	0.757	0.103	0.234	-0.281	1.279
	(0.557)	(0.793)	(0.586)	(1.105)	(0.683)	(1.590)
Math (overall)	0.572	0.663	-0.446	0.370	-0.374	2.822
	(0.819)	(1.181)	(0.507)	(0.990)	(0.807)	(1.826)
IRT measures						
Numeracy	0.039	-0.284	-0.294	-0.117	0.093	0.732
	(0.312)	(0.440)	(0.190)	(0.414)	(0.220)	(0.487)
Geometry, Algebra, Measurement	0.179	0.342	-0.223	0.043	-0.170	0.134
	(0.271)	(0.426)	(0.182)	(0.378)	(0.214)	(0.506)
Math (overall)	0.149	0.050	-0.305	-0.037	-0.058	0.551
	(0.321)	(0.491)	(0.204)	(0.434)	(0.249)	(0.563)

Note. Standard errors in parentheses. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$

Table 2.5 Robustness Check Language Models with Dichotomized Variables

	PK	**K**	**1st Grade**
Standards-Based Measures			
L&R Comprehension	0.186	-1.632	0.224
	(2.032)	(1.849)	(1.502)
Literacy Foundational Skills	0.765	-1.777	-3.149*
	(2.084)	(1.592)	(1.462)
Oral and Written Composition	-5.167	-1.797	3.546*
	(4.195)	(1.815)	(1.791)
Language/Literacy (overall)	1.407	-0.731	0.398
	(2.008)	(1.726)	(1.504)
IRT Measures			
L&R Comprehension	-1.319	-0.978	-0.038
	(2.050)	(1.568)	(1.729)
Literacy Foundational Skills	0.364	-2.086	0.285
	(1.989)	(1.515)	(1.666)
Oral and Written Composition	-0.454	-0.552	0.217
	(2.397)	(1.634)	(2.012)
Language/Literacy (overall)	0.611	-0.206	1.473
	(3.105)	(1.494)	(1.897)

Note. Standard errors in parentheses. ****p* < 0.001, ***p* < 0.01, **p* < 0.05

Table 2.6 Robustness Check Math Models with Dichotomized Variables

	PK		**K**		**1st Grade**	
	WJ-AP	**REMA**	**WJ-AP**	**REMA**	**WJ-AP**	**REMA**
Standards-based measures						
Numeracy	-0.193	-0.647	-0.522	-0.439	-0.502	1.176
	(0.486)	(0.708)	(0.308)	(0.609)	(0.351)	(0.758)
Geometry, Algebra, Measurement	-0.077	0.313	0.012	0.478	-0.370	0.213
	(0.487)	(0.711)	(0.300)	(0.587)	(0.344)	(0.748)
Math (overall)	-0.044	0.059	-0.265	0.161	-0.018	1.096
	(0.448)	(0.675)	(0.295)	(0.584)	(0.371)	(0.807)
IRT measures						
Numeracy	-0.193	-0.647	-0.634*	-0.403	0.374	0.579
	(0.486)	(0.708)	(0.294)	(0.579)	(0.671)	(1.459)
Geometry, Algebra, Measurement	0.166	0.447	-0.093	0.187	-0.403	-0.114
	(0.422)	(0.698)	(0.284)	(0.555)	(0.353)	(0.769)
Math (overall)	-0.014	-0.570	-0.398	0.514	-0.125	0.337
	(0.463)	(0.716)	(0.282)	(0.554)	(0.485)	(1.058)

Note. Standard errors in parentheses. ****p* < 0.001, ***p* < 0.01, **p* < 0.05

Chapter 3 - Study 3: Appliers to Mixed-Delivery Universal Prekindergarten Differ from Non-Appliers in Subsidy Receipt and QRIS Participation: Evidence from the Boston's UPK Expansion

Most U.S. public prekindergarten systems use a mixed-delivery approach, offering seats to children in classrooms located in public schools and in community-based organizations (CBOs) (Friedman-Krauss et al., 2021). This approach gives families more options and can help localities expand public prekindergarten more quickly. However, little is known about why some organizations would decide to participate in these mixed-delivery systems and whether are systematic differences with the organizations select that do not opt in. CBOs typically serve a population of families and children different in income, race, and ethnicity from their school-based counterparts (Crosnoe et al., 2016; Sandstrom & Chaudry, 2012; Schumacher et al., 2007; Shapiro et al., 2019) with many CBOs serving areas of concentrated disadvantage. A better understanding of how CBOs that apply to receive Universal Prekindergarten (UPK) supports differ from those that do not (non-appliers) could hold potential for improving an equitable expansion of UPK programs by identifying barriers to participation for non-appliers and by providing more detailed information on the potential supply of providers in UPK systems.

Moreover, research has shown that persistent gaps in children's outcomes by family income and race/ethnicity are present on the first day of kindergarten and that access to high-quality early education is one of the most promising routes for disrupting those early opportunity gaps and promoting equity (Chaudry et al., 2021). Implementing Universal Prekindergarten Programs (UPK) is a potential mechanism to that end. Yet, there is no consensus in the early education field on how to best achieve and maintain quality as programs go to scale. Moreover, research has shown that there can be unintended consequences of implementing UPK programs that forgo quality to increase access (Lipsey et al., 2018; van Huizen & Plantenga, 2018), which makes quality a priority for access expansion initiatives. Some empirical constraints such as the lack of population level quality assessments, population level reliable information on CBO's market prices, and potential geo-spatial patterns associated with communities' income, race, ethnicity, and additional language needs, highlight the importance of ensuring that UPK programs identify any non-intended selection patterns in their expansion phase.

To help address this gap in the literature, we leverage administrative data from the Licensing Education Analytic Database (LEAD), ratings from the Massachusetts Quality Rating and Improvement System (QRIS), subsidy records from the Child Care Financial Assistance (CCFA) system, and administrative data from the Boston Universal Prekindergarten (UPK) program to examine differences between CBOs that opted to apply to become UPK centers and those that did not. Specifically, we examine differences in centers' prekindergarten capacity, structural quality, and in the demographic characteristics of the communities and children served by the center. Boston is an excellent setting for our study because the city has been scaling out its nationally known public-school Pre-K model to CBOs since 2019 (Guerrero-Rosada, Weiland, Taylor, et al., 2021). Our findings can serve as a case study that highlights the need for systematic research on the CBOs that select into UPK systems versus those that do not in other contexts.

Center Selection into Mixed-Delivery Systems

Although mixed-delivery prekindergarten programs are a popular approach for meeting families' varied needs, these CBOs (including non-profit, Head Start, and for-profit centers) present several important equity issues systems must balance. For example, CBO teachers and administrators are often paid far less than their public-school counterparts and have lower educational levels (Garver et al., 2023). When there are setting-level differences in learning opportunities in the classroom and children's early learning gains, these tend to favor public schools (McCormick et al., 2022; Peisner-Feinberg et al., 2019; Weiland et al., 2023). Children from families with lower incomes and from minoritized racial/ethnic groups also disproportionately select into center- versus school-based programs within mixed-delivery systems (Garver et al., 2023). Accordingly, understanding how CBOs select into mixed-delivery systems and differences between appliers and non-appliers can hold promise for addressing issues of equity within these existing systems.

The present analysis is grounded in two key challenges that UPK programs face when selecting centers to participate in mixed-delivery systems. The first challenge is that UPK systems need to attract CBOs already offering high-quality services, or with the potential to achieve (with supports) high quality in the short term. Prekindergarten programs, especially those with high quality, tend to be in high demand and to have higher operational costs (Barnett & Yarosz, 2007), meaning they may be more difficult to incentivize to join UPK systems. Participating in a UPK program generally brings an additional administrative burden to programs that they would not otherwise face, just as participating in childcare subsidy systems do. As such, for CBOs, the benefits of participating must outweigh the costs.

The second challenge is that the UPK program must be equipped to identify higher-quality CBOs across the city to ensure all communities have high-quality options to meet their needs. However, at present, quality assessments are not available for the population of centers. Localities are generally faced with creating their own criteria based on information provided by the subset of applicants or collect their own information after centers have already applied. Due to these data constraints, UPK programs have a limited picture of the services available for children in some communities. A possible solution we turn to next is using *administrative data* commonly available to UPK programs to examine differences in which centers are enticed to apply versus not to join the UPK system.

Administrative Data on CBOs

Common administrative data sources for early childhood settings in the U.S. include licensing data, subsidy data for centers accepting subsidies to ease child care costs for families, and quality rating and improvement systems data. All 50 states have a licensing process for center-based preschool programs (Votruba-Drzal & Dearing, 2017) and all states have subsidy systems tied to the federal Child Care and Development Block Grant (CCDBG) (Lynch, 2022). Forty-two states have a QRIS (The Build Initiative, 2023). Further, data on community characteristics are publicly available from the American Community Survey 5-year estimates (Berkley, 2017). These data could be harnessed to support selection of centers into mixed-delivery systems and to understand features of applicant versus non-applicant centers. Below, we review research on *key indicators* from each of these systems.

Capacity

One key variable available in licensing data systems is *capacity*, meaning how many preschool-aged children is the center approved to serve. Capacity is a key center feature for a very practical reason – increasing access to preschool means offering more seats, a task made easier in centers with more space and staff available. There are also economies of scale for UPK centers to consider. Recognizing that CBOs have smaller economies of scale than public schools, for example, New Jersey pays a higher rate for slots in community-based preschool programs than for public school programs in their mixed-delivery system (Garver et al., 2023). Although empirical evidence on the association between centers' capacity and quality is scarce, a program's licensed capacity depends on their physical facilities, administrative resources to hire and retain staff, and financial resources that determine the size of centers' operation, among other aspects of structural quality.

Subsidies

Child care subsidy data systems can provide data on whether a CBO accepts any subsidies to enroll preschool-aged children and if so, the number of children gaining access to the program using a subsidy. These data may be valuable to a UPK program for several reasons. The first is that subsidy receipt may be a quality signal. Centers that enroll children with subsidies have lower quality ratings compared to centers that do not accept subsidies. (Jones-Branch et al., 2004). These findings, however, may be biased due to potential associations between neighborhoods' affluence and centers' quality. The authors posit that participation rates of centers receiving and not receiving subsidies are comparable across the county where this research took place but did not incorporate neighborhood controls in their analytical approach. The second reason is that subsidy receipt can also provide a signal that the center serves economically marginalized children and families that a locality may want to prioritize in terms of equity of access to UPK. Third, accepting subsidies is also an indicator that the center has capacity to manage different funding streams, which could be an important structural feature since UPK would add to the administrative load faced by a participating center.

Licensing Standards

States have made significant investments in licensing standards and systems for ensuring safe environments for young children. To be licensed, centers generally need to meet a set of standards on their physical environment, administration, operations, personnel, and community engagement (Gallagher et al., 1999). These too are features tracked in state administrative data that could be leveraged in mixed-delivery systems, leveraging the fact that all centers are regularly assessed for compliance with the state standards. Empirical research on licensing standards shows that more stringent state regulations increase the quality of services (Gallagher et al., 1999; Hotz & Xiao, 2011; National Association for the Education of Young Children (NAEYC), 2010), especially in higher-income areas (Hotz & Xiao, 2011). Additionally, UPK programs can leverage centers' licensing data to identify areas of low compliance and inform strategies to increase availability of high-quality seats in particular communities.

Quality Rating and Improvement Systems

As mentioned earlier, 42 states have quality rating and improvement systems, which are meant to incentivize programs to improve their quality, often via financial incentives for programs with higher quality (Thomson et al., 2020; Tout et al., 2009). In all, about one-third of centers in the U.S. were participating in QRIS in 2012 (Jenkins et al., 2021). There is a more extensive literature on QRIS systems than other administrative data typically available on centers. Notably, participating in the QRIS accountability system does seem to lead to improvements in centers' process quality scores in some cases (Bassok et al., 2019). In others, scores appear to be increased via improvements in structural characteristics like child and health screenings and director qualifications (Gomez et al., 2022). There is no consistent evidence, however, that participating in QRIS improves children's outcomes (Tout, 2013). Further, participation in QRIS appears to be higher among centers that blend funding, are accredited by the National Association for the Education of Young Children (NAEYC), and that serve communities with high poverty rates and lower proportions of Black residents (Jenkins et al., 2021). Whether a center participates in QRIS and if so, the center's rating level are data UPK programs could access, though the mixed evidence and differential engagement across communities means these data should be handled carefully.

Community and Child Characteristics

Community characteristics are another potentially important piece of data to consider in UPK systems. Prior research has shown that center-based prekindergarten classrooms serving lower-income and high-minority communities on average are rated as having lower process than center-based classrooms serving more affluent communities (Bassok & Galdo, 2016). In particular, CLASS emotional support and instructional support scores were 0.25 points lower (on a scale of 1 to 7) in communities in the highest quartile of percent poverty compared with communities in the lowest quartile. Conversely, centers located in the poorest communities employed teachers with approximately two more years of experience than centers in more affluent communities, and communities with higher proportion of Black residents, centers had lower child-to-adult ratios than centers in other communities. This is consistent with evidence from the New York UPK program, centers serving majority Black children scored 0.51 *SD* lower on the ECERS than providers serving majority White children, even among providers located within the same census tract. These differences were small and non-statistically significant when comparing centers serving majority White children to centers serving majority Hispanic and Asian students (Latham et al., 2021). UPK programs can and do use community characteristics to decide which communities to target in prekindergarten expansion. For example, Chicago had success in increasing equity of access in its expansion via prioritizing neighborhoods with lower incomes and higher unemployment (Ehrlich et al., 2020). DC similarly prioritized neighborhoods (wards) with lower incomes in its rollout of its three-year-old UPK program (Greenberg et al., 2020).

Center's Location

Geospatial analyses applied to the education field are a potential and actionable approach to depict variation at different clustering levels (e.g., census tracts, neighborhoods, zip codes) above and beyond average variation among groups (Cobb, 2020). A particular application of interest to us is using spatial tools to identify whether quality indicators show statistically significant levels of clustering at low- and high-levels of quality. For example, Schultz (2014) analyzed 199 public elementary schools in St. Louis to identify clustering of high-quality teachers, and found that highly qualified teachers were clustered in schools located in neighborhoods with lower levels of concentrated poverty and students of color. UPK programs can use a geospatial approach to identify and address clusters of quality and access disparities.

Mixed-Delivery UPK Expansion in Boston

The Boston UPK program began in 2005, offering free public prekindergarten to four-year-old children regardless of their background characteristics in school-based settings. Research has shown it has unusually high instructional quality and positive impacts on children's students' math, language, literacy, executive function, and socio-emotional skills at kindergarten entry (Chaudry et al., 2021; Weiland & Yoshikawa, 2013). In 2012, the Boston Public Schools began a pilot initiative with 11 CBOs partners that served mostly Black and Hispanic students. In April 2019, the program scaled out to additional CBOs and increased capacity to serve all age-eligible students, thus making Boston UPK a mixed-delivery system. The Boston UPK vision is to ensure equitable access to a free school day (6.5 hours per day / 180 days per year) in classrooms with an adequate teacher to child ratio (maximum 2:22 in school-based classrooms and 2:20 in center-based classrooms), in safe and age-appropriate environments. Centers are supported to offer comprehensive health and family engagement services and to sustain or adopt high-quality practices including the implementation of the *Focus on Early Learning* curriculum (Bardige et al., 2019). The Boston UPK program also offered centers a substantial pay boost for UPK teachers that placed them at the starting point of the BPS teacher pay scale (Guerrero-Rosada, Weiland, Taylor, et al., 2021). UPK centers with funded seats received about $11,000 per seat in the first year (2019–2020).

In our study years, a call for centers to participate in the UPK program was disseminated each year through several mechanisms, including the Boston Department of Early Education social networks, website, and via email to all potential applicants, namely licensed centers in the Boston area. To apply for the first two years of Boston UPK (2019 and 2020), centers needed evidence that they were a state licensed program with a physical location and capacity to serve eligible children in a four-year-old-only classroom[2]. The application also required centers to submit information about their organizational capacity and business model, financial documentation, enrollment history, staff processes and supports, and ability to align with the Boston UPK quality requirements. During the application process, center leaders and staff were invited to Q&A sessions where formal expectations for centers were shared in written materials and discussed with attendants. In addition to implementing the components of the Boston UPK program, some of these expectations included solving any licensing non-compliance issues during the first year of program participation and working towards obtaining a 3+ level in the Massachusetts QRIS (i.e., attaining moderate levels of quality as measured by self-assessments

[2] This last requirement changed for the third year of the program in 2021, when centers could apply to serve prekindergarten seats in mixed-age classrooms for three- and four-year old children (we only include the first two years of implementation in the current study).

and vetted by a technical visit, among other criteria) and being NAEYC accredited before finishing their first funding cycle by their third year as partner providers.

Center applications were assessed by the Boston UPK team to verify minimal requirements and schedule a needs assessment. The assessment served to identify center-level scopes of work and assign different levels of funding support. Some of the commitments required for Boston UPK participation were to ensure lead teachers had at least a Bachelor’s degree, that leadership and instructional staff participate in ongoing professional development and coaching on curriculum, financial management, comprehensive services, and family engagement, technology, and use of data to inform instruction. The program opened additional UPK seats and partnered with new centers in 2020 and 2021 to meet the demand for high-quality prekindergarten for all Boston families who would like a seat for their four-year-old child.

Present Study

The contribution of this paper to the literature on implementation of universal prekindergarten programs is twofold. First, we add to the current literature on equitable access to early education by identifying whether there are systematic differences in centers that apply to partner with UPK, among the population of licensed centers and among the subset of centers receiving subsidies. Second, we use geographical information systems (GIS) to explore variation in quality across neighborhoods and census block groups among the population of potential appliers to Boston UPK —namely, licensed CBOs. Specifically, we examine the following research questions:

1. Do community-based organizations applying to Boston UPK differ from non-appliers in terms of their capacity, structural quality, and the demographic characteristics of the communities where they are located?

2. Among centers receiving subsidies, do Boston UPK appliers differ from non-appliers in their capacity, structural quality, and the demographic characteristics of the children they serve?
3. Do proxies of structural quality from Boston community-based centers vary across census block groups and neighborhoods?

Method

Participants and Setting

Our sample includes the total population of Boston licensed early care and education centers in the 2018–2019 school year (*N* = 223). Of these centers, 32 applied for Boston UPK supports across the 2019–2020 (UPK year 1; *N* = 28 centers) and 2020–2021 (UPK year 2; *N* = 4 additional centers) school years. We excluded from analyses two providers in private school-based programs that applied to receive Boston UPK funding because we could not access their administrative records.

Procedures

The Institutional Review Boards at the lead and partner organizations for this study approved the human subjects plan before the commencement of study activities and the secondary data analysis. We used administrative data from the first two years of the Boston UPK program (2019 and 2020) and the Licensing Education Analytic Database (LEAD) data for the 2018–2019 school year. Additionally, we requested item-level data from licensing visits provided by the Massachusetts Department of Early Education and Care, including visits conducted between 2017 and 2022. To obtain demographic information about the communities and children served by centers, we accessed public data from the American Community Survey

5-year estimates 2019 at the census block group level and data at the child-level from the Massachusetts Child Care Financial Assistance (CCFA) system.

Measures

UPK Application Status

We used Boston UPK administrative data to create a binary indicator for whether the center applied to UPK versus did not.

Center Addresses and Licensed Capacity

We obtained center addresses, total approved number of seats for each age level (i.e., infants, toddlers, prekindergarten, and prekindergarten in mixed-age classrooms), and total seats licensed to the center for the 2018–2019 school year from the LEAD data.

Center Subsidy Receipt Status

We used data from the CCFA system to construct a binary indicator of whether the center received subsidies for at least one enrolled child or not.

Center Structural Quality

QRIS Participation and Rating. We also include an indicator of whether the center participated in the Massachusetts QRIS and whether it was rated at the three or four level, across four quality levels currently in the system. Attaining one or two stars in the MA QRIS signals that centers require quality improvement in all assessed areas, including curriculum and learning, learning environments, workforce development and qualifications, family and community engagement, leadership, and management (Executive Office of Education-Massachusetts, 2020). See Appendix E for details on the application, scoring process, and level requirements. In short, Level 1 has similar requirements to the state’s licensing process, Level 2 adds self-assessments of quality using standardized observational instruments, Level 3 requires centers to attain adequate levels of quality in self-assessed observational instruments and receive specialized technical visits from the Massachusetts DEC staff, and Level 4 requires centers to obtain adequate levels of quality as assessed by reliable observers in addition to receiving technical visits and demonstrate extensive documentation. For centers in the subsidy system, CCFA calculates up to an 8.5% add-on rate to the standard daily base rate (8.5% of current rate multiplied by the number of days) per child, based on QRIS scores.

Compliance with Licensing Standards. We identified centers' percentage of compliance with the Massachusetts Department of Early Education and Care (EEC) Licensing Standards, including Administration, Interactions Among Adults and Children, Curriculum and Progress Reports, Physical Facility Requirements, Family Involvement, Educator Qualifications and Professional Development, Ratios, Group Sizes and Supervision of Children, Health and Safety, Nutrition and Food Service, and Transportation (Franklin et al., 2003). These regulations apply to all programs providing non-residential services to children younger than 14 years old, regardless of the care setting and the ages of the children served. Programs receive scheduled visits to determine their level of compliance with regulations after they submit extensive documentation demonstrating their programs meet current regulations *(Child Care Program Licensing-Mass.Gov,* n.d.). We describe the assessment process and measure construction in Appendix E. We used data from each center's last assessment visit between April 18, 2017, and July 28, 2022, to calculate their compliance for each factor. Then, we aggregated across factors to obtain the centers' average compliance. A limitation of our measure of compliance is that spans after centers' application to Boston UPK in 2019. We are in the process of requesting historical licensing visits prior to the implementation of Boston UPK. See Appendix F, Figure

6F, for the distribution of this measure for the full population of centers and the subset of centers receiving subsidies.

Demographic Characteristics of Children and Communities Served by the Center

Characteristics of Children Receiving Subsidies. We used data from the CCFA system to identify the characteristics of children receiving subsidies served by each center, including children's subsidy eligibility factors (e.g., income, transitional assistance, housing), age, race and ethnicity, primary language spoken at home, eligibility for transportation, homeless status, and family monthly income.

Demographic Characteristics of the Community at the Center Location. We used data from the 5-year estimates of the American Community Survey (ACS) 2019. Specifically, we obtained block groups counts of children younger than five years, estimate median income in dollars amount for the last 12 months, race composition (i.e., percent of African American, Asian or Asian American, White, and Other / Two or more races population), ethnicity (percent of Hispanic or Latino population), percent of the population speaking a language other than English at home, and percent of the population with a bachelor's degree or higher. Because we do not have access to the demographic characteristics of enrolled children who do not receive CCFA subsidies, we used these measures to identify the demographic characteristics of the communities (i.e., census block groups) where appliers and non-appliers are located and identify potential demographic differences for the full population of centers.

Analytical Approach

To address our research question–*whether community-based preschools applying to Boston UPK differ from non-appliers in their capacity, quality, and the demographic characteristics of the children they serve*–we first estimated descriptive statistics and obtained unconditional differences for appliers and non-appliers using t-tests (see Table 3.1). Then, we estimated linear probability models following equations 1 and 2.

$$Applied_{jkz} = \beta_{jkz} + \varphi_{jkz} + \delta_{jk} + \lambda_k + \rho_k + (\varepsilon_{jkz} + \gamma_{kz} + \mu_z) \quad (1)$$

$$Applied|Subsidy_{jkz} = \beta_{jkz} + \varphi_{jkz} + \delta_{jkz} + \sigma_{jkz} + (\varepsilon_{jkz} + \gamma_{kz} + \mu_z) \quad (2)$$

where subscripts j, k, and *z* represent center, census block group, and neighborhood where the center is located, respectively. $Applied_{jkz}$ is an indicator of whether the program applied to serve as a Boston UPK center during the 2019–2020 or 2020–2021 school years. φ_{jkz} is a vector for centers' capacity to operate a classroom serving four years old exclusively and receive funding for subsidized seats. δ_{jkz} is a vector for centers' structural quality, which we proxy with centers' average percentage of compliance with licensing standards and an indicator of whether the center participates in the Massachusetts QRIS. λ_{kz}–is a vector for the demographic composition of the census block group including counts of total children under five years, population race and ethnicity with White as the reference group, median estimated income in the last year, percentage of population speaking a language other than English at home, and percentage of population who completed a bachelor's degree or higher. We include random intercepts for census block groups (γ_{kz}) and neighborhoods (μ_z) and a residual error term for centers (ε_{jkz}).

In equation 2, we restrict our models to centers receiving subsidies to be able to estimate differences in the demographic characteristics of children served by the center (σ_{jkz}) to address our second research question. These characteristics include race and ethnicity (with White as the reference group), monthly total family income, and percentage of children speaking a language other than English at home. For all our models, we entered predictors in conceptual blocks (capacity, quality, and children demographic characteristics) to assess magnitude and statistical significance of each factor, and then we tested all factors jointly. We report the full model's taxonomy.

To answer our third research question, we aggregated centers' quality to the census block group and neighborhood levels using the Arc-GIS Pro "summarize within" feature to describe geographical variation and conducted hotspot analysis to identify statistically significant differences in a) QRIS participation; and b) compliance with standards across the city by census block groups. Although we are unable to estimate hotspot analyses by UPK applications status due to the small number of sites across the city, we descriptively indicate where UPK appliers are located as an overlapping feature.

Results

RQ1: Do Community-Based Organizations Applying to Boston UPK Differ from Non-Appliers in Terms of their Capacity, Structural Quality, and Community Demographic Characteristics?

As shown in Table 3.1, when estimating uncontrolled differences, UPK appliers have a larger total capacity (0.81 *SD*), are approximately three times more likely to receive subsidies ($p < 0.000$) and to participate in the Massachusetts QRIS ($p < 0.000$). The difference in centers' total capacity is explained by a larger number of seats for four-year-old children (0.76 *SD*, equivalent to 26 seats, $p < 0.000$), given that there are no other differences in licensed seats for younger children. Among participant centers in the Massachusetts QRIS, UPK appliers are 17 pp more likely to be rated as level 3 or 4 (the two highest levels in the system). There are no differences in centers' compliance with licensing standards, our proxy for structural quality available at the population level. When compared to all centers, Boston UPK appliers are located in communities with a larger proportion of people of color (Black *SD* = 0.50, $p < 0.05$; Other race *SD* = 0.52, *p* < 0.05), a larger proportion of people who speak a language other than English (*SD* = 0.86, *p* < 0.000), a smaller proportion of White (*SD* = -0.75, *p* < 0.000) and college-educated people (*SD* = 0.66, *p* < 0.01), and lower median income (*SD* = 0.45, *p* < 0.05) than non-appliers (see Table 3.1).

Our linear probability models show that UPK applier and non-applier centers are statistically identical in their capacity and quality when models account for the demand for early education services as proxied by the population count of children younger than five years and the demographic characteristics of the communities served by the center–namely, the center's block group (see Table 3.2). Before accounting for the center's location, UPK appliers have a similar percentage of compliance with licensing standards, a similar likelihood of participating in the Massachusetts QRIS, and a similar capacity as non-appliers. The change in magnitude and statistical significance of the coefficient representing centers' probability of receiving subsidies once we account for the demographic composition at the census block group suggests a selection pattern based on the characteristics of communities at the center location.

RQ2: Do Boston UPK Appliers Differ from Non-appliers Receiving Subsidies in their Capacity, Structural Quality, and the Demographic Characteristics of the Children they Serve?

Before estimating differences among Boston UPK appliers and non-appliers receiving subsidies, we compared recipients and non-recipients and information is presented in Appendix F, Table 6F. In short, centers receiving subsidies in Boston have similar capacity and are in census block groups with similar demographic composition—except for the percentage of habitants with a college degree or a higher level of education (b = -0.39, *p* < 0.000)—than centers not receiving subsidies. However, centers receiving subsidies were more likely to participate in QRIS (58 pp, $p < 0.000$) and were less compliant with licensing standards (-8 pp, $p < 0.05$). We return to these differences in the discussion section.

When comparisons are restricted to the subset of centers receiving subsidies to address RQ 2, UPK appliers serve a larger share of children receiving subsidies (*SD* = 1.66, $p < 0.000$), a higher proportion of children between three and four years (9.35 pp, $p < 0.05$), and more children eligible for transportation (*SD* = 0.55, $p < 0.05$) than non-appliers (see Table 3.3). There are no differences in the subsidy eligibility factors or demographic characteristics of children enrolled in applier and non-applier centers.

Once we account for associations between capacity, quality, and demographic of enrolled children in a joint model, UPK appliers and non-appliers are statistically identical in their capacity and the demographic characteristics of enrolled children with subsidies, except that appliers are 14 pp more likely ($p < 0.01$) to participate in the Massachusetts QRIS and serve a larger proportion of children from two or more races, American-Indian or Alaska Native, Hawaiian or Other Pacific Islander background (*SD* = 0.07; $p < 0.05$; see Table 3.4).

RQ 3: Do Proxies of Structural Quality of Boston Centers Vary Across Census Block Groups and Neighborhoods?

As shown in Table 3.2 and described above, UPK applier centers were more likely to participate in the Massachusetts QRIS before applying to UPK and had similar compliance with licensing standards from 2019 to 2022, compared to non-appliers. We identified geospatial patterns in both factors. Descriptively, average levels of compliance with licensing standards vary across the city in communities with appliers and non-appliers when analyzed at the census block group and neighborhood levels. In Figure 1, Panel A, the size of the circles represents the proportion of UPK centers in the census block group. We present details of neighborhood-level variation in compliance with each factor in Appendix F Figure 7F, where neighborhoods with the lightest color (i.e., Fenway, East Boston, West Roxbury) show compliance with fewer than 70% of licensing standards. We also present descriptive evidence of variation in the proportion of centers participating in the Massachusetts QRIS and rated at levels 3 and 4 in Appendix F, Figure 8F. Only four neighborhoods have rates of participation higher than 80%, which makes it difficult for universal systems to rely on QRIS ratings in their recruitment processes.

We used hotspot analyses to identify whether some areas of the city were statistically significantly different than others in QRIS participation (see Figure 1, Panel B) and compliance with quality standards (see Figure 1, Panel C). Hotspots with statistically significantly higher QRIS participation were in two neighborhoods (Roxbury and Mattapan). A hotspot with statistically significant lower QRIS participation was located across four neighborhoods (Allston, Back Bay, Beacon Hill, and West End). Given that total compliance with license standards is expected (i.e., 100%) and most centers attain more than a 90% in this measure, our measure is not discriminative at high levels of compliance. Still, it is discriminative of centers with low compliance in the full population and among UPK centers (see Appendix F, Figure 6F). A hotspot analysis of the average compliance across factors shows that centers in the East Boston area had statistically significantly lower compliance than the rest of the city, which suggests that this area needs focalized efforts to expand high-quality services.

Discussion

Despite the ubiquity of mixed-delivery prekindergarten systems, there is no research on which centers participate in these systems and which do not, nor on how localities might incorporate administrative data to inform center selection processes. We find that in the Boston context, UPK applier and non-applier centers differ substantially in the probability of receiving subsidies and of participating in QRISs but did not differ in the demographic characteristics of communities where they were located. Geospatial analyses show that QRIS participation and compliance with licensing standards varied significantly across neighborhoods, with hotspots of high participation and low quality located in two different sets of neighborhoods. Below, we detail implications of our findings for UPK systems in turn.

Differences Between Applier and Non-applier CBOs

Regarding our first research question, we found that Boston UPK appliers had similar capacity than non-appliers. UPK centers were more likely to receive subsidies and to participate in the Massachusetts QRIS, before accounting for the demographic characteristics of census block groups where centers are located. Once we control for community characteristics, the probability of receiving subsidies is no longer statistically significant, suggesting that centers with a higher likelihood of applying to Boston UPK were in census block groups where habitants are more likely to use subsidies. This finding highlights the importance of identifying regulatory differences linked to particular funding streams, including participation in local QRISs (Schumacher et al., 2001). Specifically, UPK programs in states implementing mandatory QRISs can leverage the distinct inventory of assessments to identify potential partners and monitor population level disparities as shown in the Georgia UPK program (Bassok & Galdo, 2016), whereas UPK programs in states implementing QRISs with voluntary participation may find these data less informative due to bias linked to subsidy incentives.

We compared centers by subsidy receipt status to provide additional context on the implications of this selection pattern (see Appendix F, Table 1F). We found that centers using subsidies, overall, were more likely to participate in QRIS and to be less compliant with licensing standards than centers not receiving subsidies, consistent with prior literature indicating similar differences in quality (Johnson et al., 2019; Jones-Branch et al., 2004). In Boston, participant QRIS centers were also 66 pp more likely to receive subsidies, a logical result tied to some of the monetary incentives to increase QRIS participation (Jenkins et al., 2021). More research is needed to systematically assess to what extent subsidy recipients differ from their unsubsidized counterparts in aspects of their operation that may relate to their decision about receiving subsidies, such as their financial, operational model, and administrative staff capacity (Herbst, 2023). For example, UPK programs can use this information to design or adapt the requirements to access funding. In the Boston UPK program, centers are required to report data on staff turnover, overall enrollment, and family engagement protocols which generally require qualified staff. Centers receiving subsidies and QRIS participants may be more likely to have systems in place to account for the above information.

Similarly, more research is needed on the optimal funding and reimbursement mechanisms that UPK programs can implement to attract high-quality providers who may be reluctant to engage with state subsidies and QRIS systems. In particular, UPK programs may benefit from understanding the business model of non-subsidized centers in relation to practices such as implementing an evidence-based curriculum, sustaining a professional development model with job-embedded coaching, and securing adequate working conditions for teachers—including reduced ratios, dedicated time for planning, and adequate compensation (Bassok, Magouirk, et al., 2021; Bassok, Markowitz, et al., 2021; Weiland, 2016). Understanding how to sustainably incorporate these practices into centers' operational and financial model is important so that UPK can support centers to become fully independent after their funding cycle ends.

We also used a population-level measure of compliance with quality standards as an alternative to examine quality disparities in communities where centers are less likely to engage with QRIS. By definition, analyzing compliance limits our approach to observe variation at the lower range of centers' structural characteristics–a limitation of current screening and accountability systems (Markowitz et al., 2018). However, we considered this measure informative because, unlike QRIS participation and ratings, compliance assessments are not linked to subsidy incentives and are available with no additional costs to UPK systems. These conditions motivated us to examine the potential of compliance with standards data as a quality proxy. We found that although licensing information is not discriminative enough at adequate levels of compliance with licensing standards, this measure identifies centers with *lower structural quality at the population level*. We hypothesized items for curriculum and interactions would have larger variability and better discriminative properties and aimed to weigh these factors accordingly to differentiate centers' readiness to participate in UPK programs. Although we could not meet this goal due to data properties, UPK programs could explore synergistic efforts with licensing systems to include relevant and informative indicators of instructional quality through current installed capacity in state licensing systems.

Differences Between Subsidized Applier and Non-applier CBOs

When restricting our analysis to the subset of centers receiving subsidies for our second research question, we found differences between appliers and non-appliers in QRIS participation and in proportion of children from other and two or more races. These findings suggest that QRIS ratings may conflate information about the demand for subsidized services in some communities, consistent with prior research showing that income and racial disparities are linked to QRIS participation (Gomez et al., 2022; Jenkins et al., 2021). Although descriptively Boston UPK appliers were in communities with a higher proportion of people of color, higher linguistic diversity, and lower income in comparison with non-appliers, we found no statistically significant differences between the demographic characteristics of children attending UPK applier centers compared to children attending non-applier centers. An important limitation of our data is that we do not have access to demographic information for non-subsidized children in both appliers and non-appliers. Due to not having access to the overall demographic composition of enrolled children in the centers, our data is insufficient to make inferences about the overall centers' composition. Future research will benefit from examining the demographic composition of Boston UPK appliers and non-appliers.

Variation Across Census Block Groups and Neighborhoods

Finally, regarding our third research question, we used a geospatial approach to identify areas with a higher need for funding and quality improvement support. Research has already used this tool to monitor equitable access to high-performing teachers in elementary schools (Schultz, 2014) and applications to Boston Prekindergarten at the study level (Shapiro et al., 2019). In our approach, although we did not identify differences by application status in our linear probability models, geospatial analysis showed areas with statistically significantly higher QRIS participation and statistically significantly lower compliance with licensing standards. Both results are actionable directions for UPK programs. The former can guide recruitment efforts by indicating areas of the city with higher installed capacity, as well as potential providers with operational and financial readiness to engage with blended funding streams. The latter can guide improvement efforts at scale, by closely supporting Boston UPK centers in the area. Although an important limitation of this paper is the time span of our compliance measure, which includes visits after the rollout of Boston UPK and therefore can conflated UPK supports, it is unlikely that Boston UPK supports are related with non-appliers low compliance.

In sum, this paper has three actionable main takeaways. First, centers' financial and operational models are the most important predictors of application to the Boston UPK program. More research on barriers specific to non-subsidized centers is needed to better understand their role on UPK systems. Second, our findings illuminate the importance of monitoring quality at the population-level using measures that are not linked to subsidy incentives. Third, using neighborhood-centered approaches is a promising strategy to identify and address potential quality disparities during the scale up process.

References

Barnett, W. S., & Yarosz, D. J. (2007). Who goes to preschool and why does it matter? Preschool Policy Brief. Issue 15. *National Institute for Early Education Research.*

Bassok, D., Dee, T., & Latham, S. (2017). *The Effects of Accountability Incentives in Early Childhood Education* (No. w23859; NBER Working Paper Series). National Bureau of Economic Research. https://doi.org/10.3386/w23859

Bassok, D., & Galdo, E. (2016). Inequality in Preschool Quality? Community-Level Disparities in Access to High-Quality Learning Environments. *Early Education and Development*, *27*(1), 128–144. https://doi.org/10.1080/10409289.2015.1057463

Bassok, D., Magouirk, P., & Markowitz, A. J. (2021). Systemwide Quality Improvement in Early Childhood Education: Evidence From Louisiana. *AERA Open*, *7*, 23328584211011610.

Bassok, D., Markowitz, A. J., Bellows, L., & Sadowski, K. (2021). New Evidence on Teacher Turnover in Early Childhood. *Educational Evaluation and Policy Analysis*, *43*(1), 172–180. https://doi.org/10.3102/0162373720985340

Berkley, J. (2017). Using American community survey estimates and margins of error. *United States Census Bureau*.

Chaudry, A., Morrissey, T., Weiland, C., & Yoshikawa, H. (2021). *Cradle to kindergarten: A new plan to combat inequality* (2nd edition). Russell Sage Foundation.

Child Care Program Licensing | Mass.gov. (n.d.). Retrieved April 16, 2023, from https://www.mass.gov/child-care-program-licensing

Cobb, C. D. (2020). Geospatial analysis: A new window into educational equity, access, and opportunity. *Review of Research in Education*, *44*(1), 97–129.

Crosnoe, R., Benner, A. D., & Davis-Kean, P. (2016). Preschool Enrollment, Classroom Instruction, Elementary School Context, and the Reading Achievement of Children from Low-Income Families. In G. Kao & H. Park (Eds.), *Research in the Sociology of Education* (Vol. 19, pp. 19–47). Emerald Group Publishing Limited. https://doi.org/10.1108/S1479-353920150000019003

Ehrlich, S. B., Connors, M. C., Stein, A. G., Francis, J., Easton, J. Q., Kabourek, S. E., & Farrar, I. C. (2020). Closer to home: More equitable pre-k access and enrollment in Chicago. Research report. *Start Early*.

Franklin, S. P., Lamana, A., & Van Thiel, L. (2003). *Early Childhood Program Standards for Three- and Four-Year Olds.*

Friedman-Krauss, A. H., Barnett, W. S., Garver, K. A., Hodges, K. S., Weisenfeld, G. G., & Gardiner, B. A. (2021). The state of preschool 2020: State preschool yearbook. *National Institute for Early Education Research*.

Gallagher, J. J., Rooney, R., & Campbell, S. (1999). Child care licensing regulations and child care quality in four states. *Early Childhood Research Quarterly, 14*(3), 313–333. https://doi.org/10.1016/S0885-2006(99)00015-0

Garver, K., Weisenfeld, G. G., Connors-Tadros, L., Hodges, K., Melnick, H., & Placencia, S. (2023). *State preschool in a mixed delivery system: Lessons from five states*. Learning Policy Institute. https://doi.org/10.54300/387.446

Gomez, C. J., Whitaker, A. A., & Cannon, J. S. (2022). Do early care and education programs improve when enrolled in quality rating and improvement systems? Longitudinal evidence from one system. *Early Education and Development*, 1–18. https://doi.org/10.1080/10409289.2022.2105624

Greenberg, E., Luetmer, G., Chien, C., & Monarrez, T. (2020). *Who Wins the Preschool Lottery? Applicants and Application Patterns in DC Public Prekindergarten. Research Report.* Urban Institute.

Guerrero-Rosada, P., Weiland, C., Taylor, A., Penfold, L., Snow, C. E., Sachs, J., & McCormick, M. (2021). *Effects of COVID-19 on Early Childhood Education Centers: Descriptive Evidence from Boston's Universal Prekindergarten Initiative*. Gerald R. Ford School of Public Policy, University of Michigan. Education Policy Initiative.

Herbst, C. M. (2023). Child care in the United States: Markets, policy, and evidence. *Journal of Policy Analysis and Management*, *42*(1), 255–304. https://doi.org/10.1002/pam.22436

Hotz, V. J., & Xiao, M. (2011). The impact of regulations on the supply and quality of care in child care markets. *American Economic Review*, *101*(5), 1775–1805. https://doi.org/10.1257/aer.101.5.1775

Jenkins, J. M., Duer, J. K., & Connors, M. (2021). Who participates in quality rating and improvement systems? *Early Childhood Research Quarterly*, *54*, 219–227. https://doi.org/10.1016/j.ecresq.2020.09.005

Johnson, A. D., Martin, A., & Schochet, O. N. (2019). How do early care and education workforce and classroom characteristics differ between subsidized centers and available center-based alternatives for low-income children? *Children and Youth Services Review*, *107*, 104567. https://doi.org/10.1016/j.childyouth.2019.104567

Jones-Branch, J. A., Torquati, J. C., Raikes, H., & Pope Edwards, C. (2004). Child care subsidy and quality. *Early Education & Development*, *15*(3), 327–342. https://doi.org/10.1207/s15566935eed1503_5

Latham, S., Corcoran, S. P., Sattin-Bajaj, C., & Jennings, J. L. (2021). Racial disparities in pre-k quality: Evidence from New York City’s universal pre-k program. *Educational Researcher*, 0013189X2110282. https://doi.org/10.3102/0013189X211028214

Learn about the Massachusetts Quality Rating and Improvement System (QRIS) | Mass.gov. (n.d.). Retrieved April 16, 2023, from https://www.mass.gov/service-details/learn-about-the-massachusetts-quality-rating-and-improvement-system-qris

Lipsey, M. W., Farran, D. C., & Durkin, K. (2018). Effects of the Tennessee Prekindergarten Program on children’s achievement and behavior through third grade. *Early Childhood Research Quarterly, 45*, 155-176.

Lynch, K. E. (2022). The Child Care and Development Block Grant: In Brief. CRS Report R47312, Version 4. *Congressional Research Service*.

Markowitz, A. J., Bassok, D., & Hamre, B. (2018). Leveraging Developmental Insights to Improve Early Childhood Education. *Child Development Perspectives*, *12*(2), 87–92. https://doi.org/10.1111/cdep.12266

National Association for the Education of Young Children (NAEYC). (2010). *Best Practices of Accreditation Facilitation Projects: A framework for Program Quality Improvement Using NAEYC Early Childhood Program Standards and Accreditation Criteria*. NAEYC. https://www.naeyc.org/sites/default/files/globally-shared/downloads/PDFs/accreditation/early-learning/AFPBestPractices.pdf

Sandstrom, H., & Chaudry, A. (2012). ‘You have to choose your childcare to fit your work’: Childcare decision-making among low-income working families. *Journal of Children and Poverty, 18*(2), 89-119.

Schultz, L. M. (2014). Inequitable dispersion: Mapping the distribution of highly qualified teachers in St. Louis metropolitan elementary schools. *Education Policy Analysis Archives*, *22*(90), n90.

Schumacher, R., Greenberg, M., & Lombardi, J. (2001). *State Initiatives To Promote Early Learning: Next Steps in Coordinating Subsidized Child Care, Head Start, and State Prekindergarten. Full Report.*

Shapiro, A., Martin, E., Weiland, C., & Unterman, R. (2019). If you offer it, will they come? Patterns of application and enrollment behavior in a universal prekindergarten context. *AERA Open*, *5*(2), 2332858419848442.

The Build Initiative. (2023). *Quality Rating and Improvement Systems Compendium*. https://qualitycompendium.org/create-a-report

Thomson, D., Cantrell, E., Guerra, G., Gooze, R., & Tout, K. (2020). *Conceptualizing and measuring access to early care and education*.

Tout, K. (2013). Look to the stars: Future directions for the evaluation of quality rating and improvement systems. *Early Education & Development*, *24*(1), 71–78. https://doi.org/10.1080/10409289.2013.741912

Tout, K., Zaslow, M., Halle, T., & Forry, N. (2009). *Issues for the next decade of quality rating and improvement systems* (No. 2009-14; Issue Brief). Child Trends. https://www.acf.hhs.gov/sites/default/files/opre/next_decade.pdf

van Huizen, T., & Plantenga, J. (2018). Do children benefit from universal early childhood education and care? A meta-analysis of evidence from natural experiments. *Economics of Education Review*, *66*, 206-222.

Votruba-Drzal, E., & Dearing, E. (Eds.). (2017). *The Wiley handbook of early childhood development programs, practices, and policies*. Wiley Blackwell.

Weiland, C. (2016). Launching Preschool 2.0: A road map to high-quality public programs at scale. *Behavioral Science & Policy, 2*(1), 37–46. https://doi.org/10.1353/bsp.2016.0005

Weiland, C., & Yoshikawa, H. (2013). Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *Child Development*, *84*(6), 2112–2130. https://doi.org/10.1111/cdev.12099

Table 3.1 Centers' Baseline Capacity, Quality, and Demographics at their Location by Application Status

	Non-appliers (*N* = 191)		**UPK appliers (*N* = 32)**		
	Mean or %	***SD***	**Mean or %**	***SD***	**Difference**
Capacity					
Total licensed capacity	57.00	41.14	90.34	51.75	33.34***
Infant (birth – 15 months)	5.06	7.21	6.12	7.90	1.05
Toddler (15 months – 33 months)	9.16	12.78	10.55	12.20	1.39
PreK (33 months – Kindergarten)	32.88	33.48	59.21	34.85	26.33***
PreK in mixed-age classrooms	2.02	6.94	1.13	6.36	-0.87
Receives EEC subsidies	52.36	--	87.50	--	35.14***
Quality					
In QRIS	51.83	--	87.50	--	35.67***
QRIS 3+ (127 QRIS participants)	8.08	--	25.00	--	16.91*
Average Licensing Compliance	94.36	5.54	95.11	4.04	0.07
Administration	89.02	11.38	88.35	12.20	-0.07
Staff and Ratios	95.98	13.07	99.42	3.09	0.34
Facilities	93.43	12.25	94.79	9.66	1.36
Health and Safety	85.75	16.03	85.71	14.03	-0.00
Nutrition	98.39	6.53	98.21	9.45	-0.02
Interactions	98.98	7.96	99.42	3.09	0.04
Curriculum	99.46	4.39	100.00	0.00	0.05
Demographics at the centers' location					
Children under five years	68.97	61.84	88.29	83.87	19.31
% Black or African American	22.36	26.94	36.49	29.36	14.13*
% Asian or Asian American	10.63	12.25	14.20	21.09	3.56
% Other or Mixed	12.12	11.00	15.92	12.07	3.80
% Hispanic or Latino	19.32	17.24	24.42	19.18	5.10
% White	54.88	27.17	33.39	28.22	-21.50***
Median Income Dollars	82,238.50	46,611.78	59,707.58	47,411.48	-22,530.91*
% Speak a Language other than English	36.03	16.67	50.19	21.77	14.16***
% College Degree +	52.55	26.47	34.68	21.94	-17.87**

Note. ****p* < 0.001, ***p* < 0.01, **p* < 0.05. We excluded two licensing factors only assessed for a small number of centers (i.e., transportation *N* = 117 and family involvement *N* = 56). UPK non-appliers are distributed across 173 census block groups, and appliers are distributed across 173. Only ten block groups (out of 201 in Boston) have both appliers and non-appliers.

Table 3.2 Taxonomy of Linear Probability Models Predicting Application to Boston UPK – Full Population

	Center Applied to Boston UPK			
	(1)	**(2)**	**(3)**	**(4)**
Center's capacity				
Total PreK Capacity	0.00*			0.00
	(0.00)			(0.00)
Receives Subsidies	0.13**			0.05
	(0.04)			(0.04)
Proxies of Structural Quality				
Participates in QRIS		0.14***		0.03
		(0.04)		(0.03)
Average Compliance with Standards		0.12		0.15
		(0.26)		(0.22)
Community characteristics at the center location				
Children under 5 years old			0.00	0.00
			(0.00)	(0.00)
% Asian			0.00	0.00
			(0.00)	(0.00)
% Black or African American			0.00	0.00
			(0.00)	(0.00)
% Hispanic or Latino			-0.00	-0.00
			(0.00)	(0.00)
% Other and Mixed			0.00	0.00
			(0.00)	(0.00)
Estimate Median household income in the past 12 months			0.00	0.00
			(0.00)	(0.00)
% Speak other languages			0.00	0.00
			(0.00)	(0.00)
% Bachelor's degree or higher			-0.00	-0.00
			(0.00)	(0.00)
Constant	-0.04	-0.06	0.06	-0.27
	(0.06)	(0.25)	(0.15)	(0.22)
Observations	193	193	193	193
Neighborhoods	16	16	16	16

Note. ***p < 0.001, **p < 0.01, *p < 0.05 Robust standard errors in parentheses

Table 3.3 Characteristics of Children Served by Non-UPK and UPK Centers During the 2018–2019 School Year, in Centers Receiving Subsidies

	Non-UPK (*N* = 100)		**UPK (*N* = 28)**		**Difference**
	Mean or %	***SD***	**Mean or %**	***SD***	
Children served by centers					
Enrolled children receiving subsidies	27.58	23.30	66.41	46.87	38.83***
Enrolled children eligible for transportation	7.44	13.47	14.89	25.88	7.45*
Enrolled children with homeless status	0.84	4.14	2.53	7.02	1.69
Children’s age by September 1, 2018	3.98	1.97	3.57	1.31	-0.41
% Children under one year	5.99	9.93	6.45	6.60	0.46
% Children between 1 and 2 years	14.00	15.09	13.23	13.23	-0.76
% Children between 2 and 3 years	18.85	15.19	19.83	9.25	0.09
% Children between 3 and 4 years	20.59	17.21	29.94	15.32	9.35*
% Children between 4 and 5 years	16.12	16.96	16.27	10.57	0.14
Female	48.47	16.42	51.09	6.52	2.61
Subsidies payments and eligibility					
Monthly total family income	2349.30	648.83	2296.87	439.40	-52.43
Total dollar amount billed by the provider	882.20	278.38	969.72	174.08	87.52
Dollar amount of subsidies received	887.48	282.90	979.59	180.50	92.10
DCF (Department of Children and Families)	12.49	17.94	9.94	9.24	-2.54
DHCD (Department of Housing and Community)	1.61	6.75	5.12	14.04	3.51
DTA (Department of Transitional Assistance)	15.69	16.18	9.99	6.77	-5.70
DTA-PT	7.92	12.38	4.53	4.53	-3.38
DTA-T	5.05	5.54	3.47	3.59	-1.15
Income Eligible	57.23	25.81	66.94	16.83	9.70
Children’s race/ethnicity					
% Asian or Asian American	16.92	13.02	15.44	9.51	-1.48
% Black or African American	29.71	29.71	29.22	11.88	0.49
% Hispanic or Latino	22.82	19.72	26.42	18.89	3.59
% Two or More Races and Other	11.05	8.56	13.06	9.46	2.00
% White	19.49	14.03	15.85	6.95	-3.63
Language spoken at home					
% Chinese	1.00	5.38	4.22	16.99	3.22
% English	84.31	18.74	79.15	20.00	-5.14
% Spanish	11.27	15.22	12.50	12.67	1.22
% Other languages	1.69	3.83	3.59	6.12	1.86~

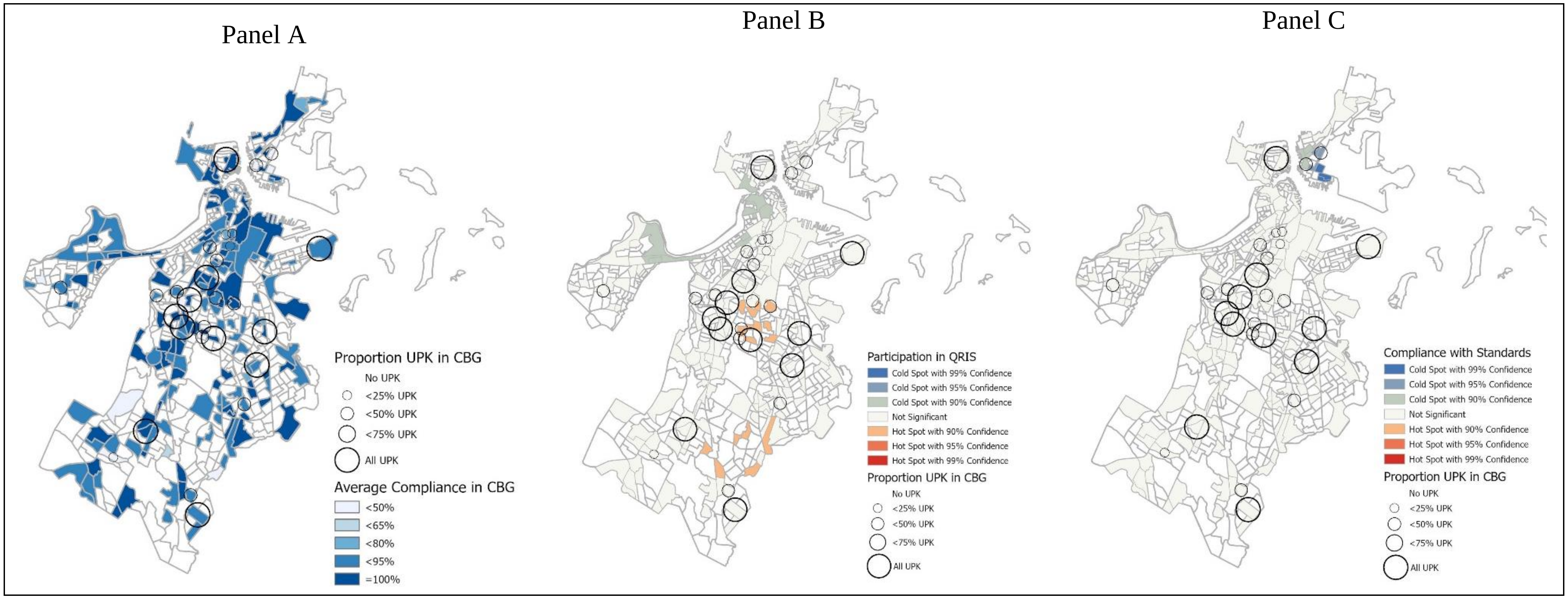
Note. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

Table 3.4 Taxonomy of Linear Probability Models Predicting Application to Boston UPK–Receiving Subsidies

	Center Applied to UPK			
	(1)	**(2)**	**(3)**	**(4)**
Center's Capacity				
Total PreK Capacity	0.00			0.00
	(0.00)			(0.00)
Proxies of Structural Quality				
Participates in QRIS		0.15*		0.12**
		(0.07)		(0.04)
Average Compliance with Standards		0.43		0.21
		(0.38)		(0.28)
Demographic characteristics of enrolled children receiving subsidies				
Family Monthly Income			-0.00	-0.00
			(0.00)	(0.00)
% Asian or Asian American			0.34	0.07
			(0.32)	(0.22)
% Black or African American			0.23	0.04
			(0.20)	(0.28)
% Hispanic or Latino/a			0.42	0.21
			(0.47)	(0.32)
% Other and Mixed			0.96***	0.63*
			(0.26)	(0.29)
% Speaks English at Home			-0.20	-0.04
			(0.25)	(0.22)
Constant	0.13	-0.27	0.20	-0.16
	(0.08)	(0.31)	(0.36)	(0.45)
Observations	122	122	122	122
Neighborhood	16	16	16	16

Note. ****p* < 0.001, ***p* < 0.01, **p* < 0.05. Robust standard errors in parentheses.

Figure 3.1 Distribution of Quality Through Two Different Measures: Participation in QRIS and Compliance with Standards


Panel A

Panel B

Panel C

Proportion UPK in CBG
No UPK
<25% UPK
<50% UPK
<75% UPK
All UPK
Average Compliance in CBG
<50%
<65%
<80%
<95%
=100%
Participation in QRIS
Cold Spot with 99% Confidence
Cold Spot with 95% Confidence
Cold Spot with 90% Confidence
Not Significant
Hot Spot with 90% Confidence
Hot Spot with 95% Confidence
Hot Spot with 99% Confidence
Proportion UPK in CBG
No UPK
<25% UPK
<50% UPK
<75% UPK
All UPK
Compliance with Standards
Cold Spot with 99% Confidence
Cold Spot with 95% Confidence
Cold Spot with 90% Confidence
Not Significant
Hot Spot with 90% Confidence
Hot Spot with 95% Confidence
Hot Spot with 99% Confidence
Proportion UPK in CBG
No UPK
<25% UPK
<50% UPK
<75% UPK
All UPK

Conclusion

All but one state delivers public prekindergarten through a mixed-delivery system with classrooms in public schools and community-based organizations (Friedman-Krauss et al., 2021). Yet, there is very little research on *how* to sustain and scale up quality across these settings. The context of my three dissertation studies was a privileged one to examine actionable features of quality that classrooms, schools, and systems can consider when working towards sustaining and expanding quality in large-scale programs. As part of the ExCEL P-3: Promoting Sustained Gains from Preschool to Third Grade (Hsueh, 2016) and the Boston Universal Pre-K (UPK; Weiland and Snow, 2019) studies, research teams collected micro-level observational data on classroom's instruction (i.e., how children spent their time during two regular days in the program), survey data across early education grades (i.e., what teachers primarily taught during prekindergarten, kindergarten, and first grade), and administrative data on the scale-up process of the Boston UPK program (i.e., what centers have historically applied during the first three years of program implementation). I used these data sources in the context of a program nationally recognized for its positive impacts on children's development to examine actionable instructional features of the program at the classroom, school, and system levels.

I approached this question through two lenses. First, a measurement consistency lens. The early education field faces increasing challenges to achieve consensus on how high-quality looks and what components predict children's gains (Burchinal, 2017; Weiland, 2018; Weiland & Guerrero-Rosada, 2022). Examining consistency across instruments is a necessary step to identify how measurement is deterring the field progress toward improving quality. Second, an equitable implementation lens. There is increasing evidence of early and persistent opportunity gaps that prekindergarten programs have the potential to address (Chaudry et al., 2021; Magnuson & Duncan, 2016). Classrooms, schools, and systems have transformed to meet the needs of more diverse populations, which calls for new analytical approaches that can accurately reflect these contexts.

I explored measurement consistency in my three studies. At the classroom level, in my time use study, I identified aspects of the coding protocols and instruments' design that introduce systematic noise to classroom measurement. At the school level, in my instructional complexity study, I found that conceptually and empirically based measures capture vertical misalignment and consistently assess the variation in content complexity across grade levels, especially for unconstrained content domains (i.e., language and comprehension and numeracy). At the system level, in my UPK appliers study, I attempted to compose a population-level quality proxy (i.e., compliance with licensing standards) but failed to do so due to a lack of variability in standards related with curriculum implementation and interactions. Results from studies 1 and 2 can be used by school districts and the early education field to better calibrate detailed observational and survey instruments, especially in the areas of content-rich instruction and vertical alignment. My third study can be used by the UPK program to monitor quality hotspots trough Boston and, potentially, create synergetic efforts to improve the assessment of instruction and curriculum standards in the licensing system.

Regarding equitable implementation, I provide evidence in study 1 that, at least in school-based prekindergarten classrooms, the type of instruction to which children are exposed varied across demographic groups. Recent research has identified this within-classroom variation (Weiland et al., 2023). However, I add evidence of similar variation across classrooms with different demographic composition, which school districts can examine further and address through professional development. Although I did not directly examine variation by demographic characteristics in my second study, the different levels of exposure to content complexity across classrooms and grade-levels suggest large differences in the instruction that children receive, which may have implications for later achievement (Maier et al., 2022). At the system level, I provide evidence that CBOs applying to the Boston UPK program serve similar communities to non-appliers. These centers, however, are statistically significantly more likely to receive subsidies and to participate in the QRIS system. This information can be used by universal prekindergarten programs to assess whether their recruitment, monitoring, and accountability processes disincentivize potential partner providers that are less likely to engage with subsidized funding streams and large-scale accountability.

Despite having good descriptive properties, the measures I explored did not consistently predict children's gains. When they did, statistically significant associations depended on the modeling approach. A potential explanation is that the low predictive power of the measures I utilized in studies 1 and 2 relates to these measures' limitations to capture the quality of instruction in specific content areas. Another potential explanation is a lack of direct correspondence between the predictors and outcomes I examined. In math and literacy, for example, authors have identified good practices that are specific to each content domain and exceed the scope of more general measures (Clements et al., 2013; Duke & Del Nero, 2011; Sarama & Clements, 2009; Snow et al., 1998). Further research can extend my findings throughout these studies to develop new content-specific quality measures informing prekindergarten, kindergarten, and first grade instruction.

References

Bardige, B. L. S., Baker, M., Mardell, B., Boston Public Schools, & Department of Early Childhood. (2018). *Children at the center: Transforming early childhood education in the Boston Public Schools*. https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=2894453

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.

Burchinal, M. (2017). Measuring early care and education quality. *Child Development Perspectives*, *12*(1), 3–9. https://doi.org/10.1111/cdep.12260

Chaudry, A., Morrissey, T., Weiland, C., & Yoshikawa, H. (2021). *Cradle to kindergarten: A new plan to combat inequality* (2nd edition). Russell Sage Foundation.

Clements, D. H., Agodini, R., & Harris, B. (2013). Instructional Practices and Student Math Achievement: Correlations from a Study of Math Curricula. NCEE Evaluation Brief. NCEE 2013-4020. *National Center for Education Evaluation and Regional Assistance*.

Duke, N. K., & Del Nero, J. R. (2011). *Best practices in literacy instruction*. Guilford Press.

Friedman-Krauss, A. H., Barnett, W. S., Garver, K. A., Hodges, K. S., Weisenfeld, G. G., & Gardiner, B. A. (2021). The state of preschool 2020: State preschool yearbook. *National Institute for Early Education Research*.

Hsueh, J. (2016). *Boston P-3: Identifying Malleable Factors for Promoting Student Success*. Institute of Education Sciences-Funded Research Grants and Contracts. https://ies.ed.gov/funding/grantsearch/details.asp?ID=1770

Magnuson, K. & Duncan, G.J. (2016). Can early childhood interventions decrease inequality of economic opportunity? *RSF: The Russell Sage Foundation Journal of the Social Sciences*, *2*(2), 123. https://doi.org/10.7758/rsf.2016.2.2.05

Maier, M. F., McCormick, M. P., Xia, S., Hsueh, J., Weiland, C., Morales, A., Boni, M., Tonachel, M., Sachs, J., & Snow, C. (2022). Content-rich instruction and cognitive demand in prek: Using systematic observations to predict child gains. *Early Childhood Research Quarterly*, *60*, 96–109. https://doi.org/10.1016/j.ecresq.2021.12.010

McCormick, M. P., Mattera, S. K., Maier, M. F., Xia, S., Jacob, R., & Morris, P. A. (2022). Different settings, different patterns of impacts: Effects of a Pre-K math intervention in a mixed-delivery system. *Early Childhood Research Quarterly, 58*, 136-154.

Peisner-Feinberg, E., Van Manen, K., Mokrova, I., & Burchinal, M. (2019). Children's Outcomes through Second Grade: Findings from Year 4 of Georgia's Pre-K Longitudinal Study. *FPG Child Development Institute*.

Sarama, J., & Clements, D. H. (2009). *Manual for classroom observation (COEMET)*. Authors.

Snow, C. E., Burns, M. S., & Griffin, P. (Eds.). (1998). Preventing reading difficulties in young children. *National Academy Press*.

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

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. *MDRC*.

Weiland, C., Moffett, L., Rosada, P. G., Weissman, A., Zhang, K., Maier, M., Snow, C., McCormick, M., Hsueh, J., & Sachs, J. (2023). Learning experiences vary across young children in the same classroom: Evidence from the individualizing student instruction measure in the Boston Public Schools. *Early Childhood Research Quarterly*, *63*, 313–326. https://doi.org/10.1016/j.ecresq.2022.11.008

Weiland, C., Unterman, R., Dynarski, S., Abenavoli, R., Bloom, H., Braga, B., ... & Weixler, L. (2023). Lottery-Based Evaluations of Early Education Programs: Opportunities and Challenges for Building the Next Generation of Evidence.

Appendices

Appendix A: Models Using Standardized Versions of Outcome Measures

Table 1A. Taxonomy of Models Using Standardized Outcome Measures

	PPVT			**WJ-AP**		
NR Models						
Balanced/Mixed	2.49	1.72	1.14	-2.54*	-2.87*	-1.52
	(1.86)	(2.04)	(2.26)	(1.24)	(1.32)	(1.45)
ISI Models						
Balanced/Mixed	-2.11	-2.85	-0.78	1.45	1.05	1.80
	(1.77)	(2.05)	(2.99)	(1.43)	(1.53)	(2.03)
Child and family demographics	X	X	X	X	X	X
Process and structural quality covariates		X	X		X	X
Classroom composition covariates			X			X

Note. The reference group is the: *Whole Group / High Academic* for both measures. Standard Errors in parenthesis. ****p* < 0.001; ***p* < 0.01; **p* < 0.05. Models include random intercepts for classrooms and schools.

Appendix B: Minutes in Learning Settings and Content Areas by Profile Membership

Table 2B. Minutes in Learning Settings and Content Areas by Profile Membership

	Narrative Record						**ISI**					
	SGC/Mixed (*N* = 3; 8.6%)		**Whole Group/ High Academic** (*N* = 19, 54.3%)		**Balanced/Mixed** (*N* = 13, 37.1%)		**Balanced/ Moderate content** (*N* = 16; 45.7%)		**Whole Group/ High Academic** (*N* = 14, 40.0%)		**High Whole Group/Mixed** (*N* = 5, 14.3%)	
	Mean	***SD***	***Mean***	***SD***	***Mean***	***SD***	***Mean***	***SD***	***Mean***	***SD***	***Mean***	***SD***
Whole Group	31.30	6.77	81.31	21.22	44.11	16.21	59.98	30.75	104.28	39.62	91.29	6.73
Centers	6.44	8.60	28.47	20.39	49.40	27.37	67.38	25.19	65.62	25.72	51.08	7.27
Small Group	0.00	0.00	6.67	9.86	0.44	1.50	10.31	12.78	16.82	17.83	4.71	6.63
SG/Centers	61.42	4.81	11.55	12.29	14.93	14.61	--	--	--	--	--	--
Individual	--	--	--	--	--	--	36.84	20.15	25.02	14.82	8.68	9.92
Transitions	16.28	3.27	29.16	9.41	42.28	9.85	38.55	12.97	33.37	10.37	34.72	5.88
Language	17.59	5.01	40.78	20.70	21.48	10.87	41.55	13.92	70.60	22.60	42.90	8.44
Math	6.87	4.44	22.52	15.41	16.23	11.10	18.68	9.44	27.82	12.09	15.97	5.69
Other content	3.19	2.76	16.26	9.42	8.13	6.87	24.74	12.58	19.24	14.32	37.54	12.36
Mixed content	71.06	17.20	48.05	21.19	63.07	32.93	70.95	35.57	77.83	32.53	98.75	11.58
Total minutes	115.44	15.57	157.22	40.50	151.17	39.12	174.44	54.70	210.63	53.00	155.62	14.32

Note. The NR differentiates the proportion of time when some children are learning in Small Group and other are learning in Centers simultaneously (SG/C).

Appendix C: Properties of Conceptually Based Measures (Grade Level of Instruction)

To obtain parameters of internal consistency of teachers' reports within each scale, we estimated Cronbach's alpha coefficients. Since we are interested on to what extent teachers' answers are correlated and all items are measured using the same scale in their original report, we obtained standardized coefficients. Additionally, we declared the items to have a positive sign to obtain comparable estimations across years, since we do not allow items to vary their sign when the item pool changes to include more difficult items from the Common Core. See the items distribution per year (#), the inter-item correlations (IIC) and Cronbach's alpha (α) in Table 1C.

The average inter-item correlations (IIC) provide information about the scale homogeneity and, importantly, is unrelated to scale length – a limitation of Cronbach's alpha, which increases with the number of items in a measure. Although there are not strict cut-offs to interpret IICs, smaller values indicate low homogeneity whereas higher values are indicative of more consistent scales (Values between 0.10 and 0.25 are usually interpreted as a homogeneous but not redundant scales). Most of the scales fall under this range, except for Language and Reading Comprehension in PreK.

Cronbach's alpha (α) provides information about the covariance and correlation among the scale's items. Although a value of 0.70 is adequate, a minimum of 0.80 is desirable. This parameter (α) is not optimal for some PreK scales (*Language and Reading Comprehension* and *Numeracy* due to a small item pool), but it improves for Kindergarten and First Grade when more items were added (and variation in teachers' responses increased).

Table 3C. Structure and Reliability of Math and Language Measures

	PreK			Kindergarten			First Grade			Across years		
	#	IIC	α	#	IIC	α	#	IIC	α	#	IIC	A
Language and Literacy	*31*	*0.09*	*0.74*	*32*	*0.13*	*0.82*	*54*	*0.13*	*0.89*	*54*	*0.12*	*0.88*
L&R Comprehension	14	0.08	0.55	14	0.16	0.73	22	0.10	0.70	22	0.13	0.76
Literacy Foundational Skills	13	0.13	0.66	11	0.17	0.69	18	0.20	0.82	18	0.16	0.78
Writing (composition)	4	0.36	0.63	7	0.25	0.67	14	0.33	0.87	14	0.29	0.84
Math	*26*	*0.10*	*0.73*	*32*	*0.10*	*0.77*	*46*	*0.17*	*0.90*	*46*	*0.07*	*0.85*
Numeracy	7	0.13	0.47	11	0.12	0.61	15	0.39	0.90	15	0.19	0.77
Operations-Geometry-Measurement	19	0.12	0.71	21	0.13	0.76	32	0.15	0.84	37	0.07	0.82

Note. In operations-geometry-measurement, there were 8 items that were asked in PK only and, therefore, needed to be dropped from the model to estimate reliability across years.

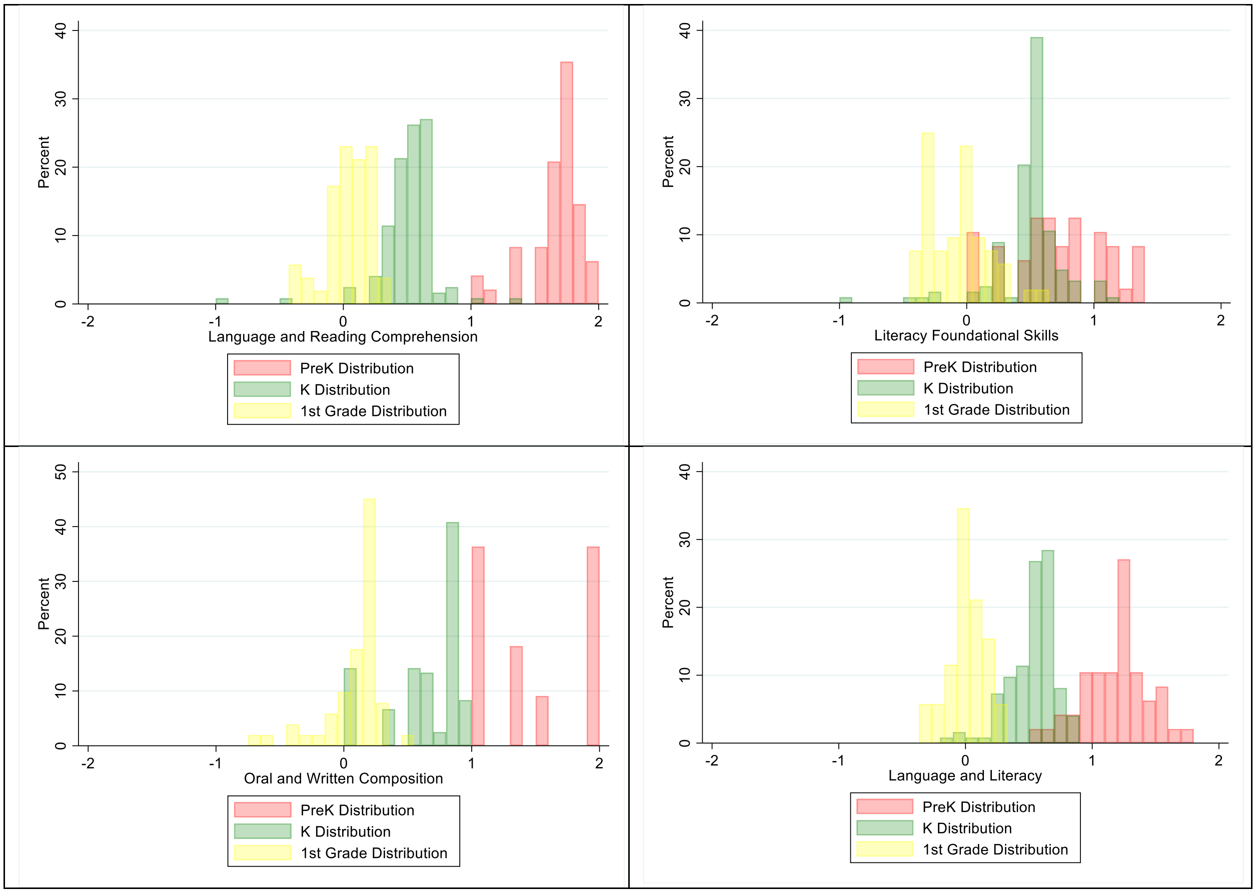


Figure 2C. Histograms of Language and Literacy Standards-Based Measures by Year

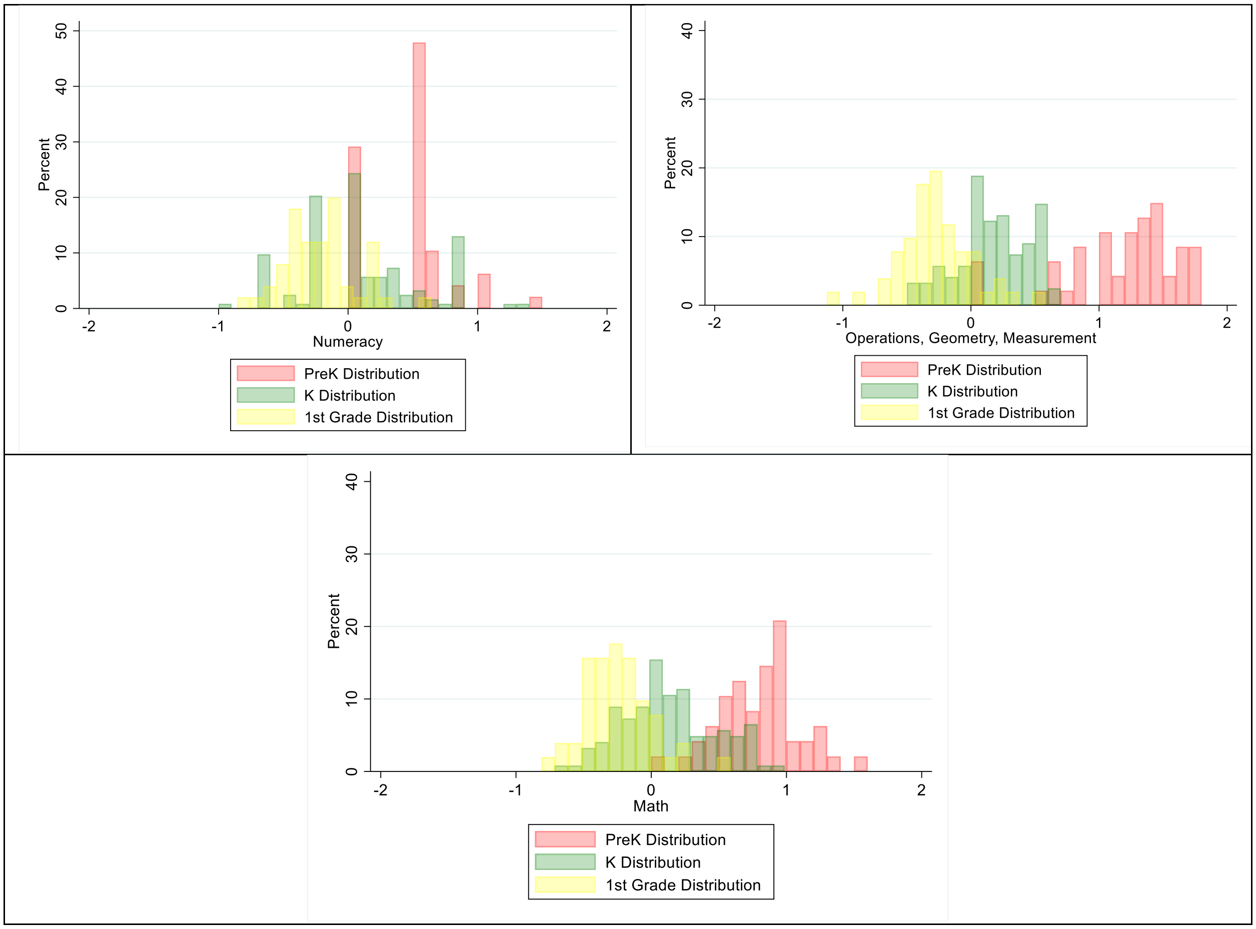


Figure 3C. Histograms of Standards-based Math Measures by Year

Appendix D: Properties of Empirically Based Measures (IRT)

We initially included all items for each sub-scale in our conceptually based measures to estimate Rasch models. To obtain unidimensional parameters, we removed items that impeded models' convergence due to high missingness or a lack of variability, evidenced by flat Item Characteristic Curves. After doing so, we estimated items' fit statistics and removed those that did not fit the model. Finally, we conducted a unidimensionality test using Modified Parallel Analysis. This test implements the procedure proposed by Drasgow and Lissak (1983) for examining the latent dimensionality of dichotomously scored item responses. In other words, it allows to assess whether the set of dichotomic items conforming each scale reliably represent one latent trait by testing whether there is a statistically significant additional parameter in the data. See Table 2D for a summary of items retained and evidence for unidimensionality of each construct of interest.

Table 4D. Structure and Reliability of Empirically Based Math and Language Measures

	Starting Items	Non-discriminative	Do not fit Rasch model	Items retained	Unidimensionality test
Language and Literacy					
L&R Comprehension	22	5	2	15	*p*-value: 0.16
Literacy Foundational Skills	18	4	3	11	*p*-value: 0.07
Writing (composition)	14	3	1	10	*p*-value: 0.24
Math					
Numeracy	15	7	0	8	*p*-value: 0.75
Operations-Geometry-Measurement	37	15	0	22	*p*-value: 0.27

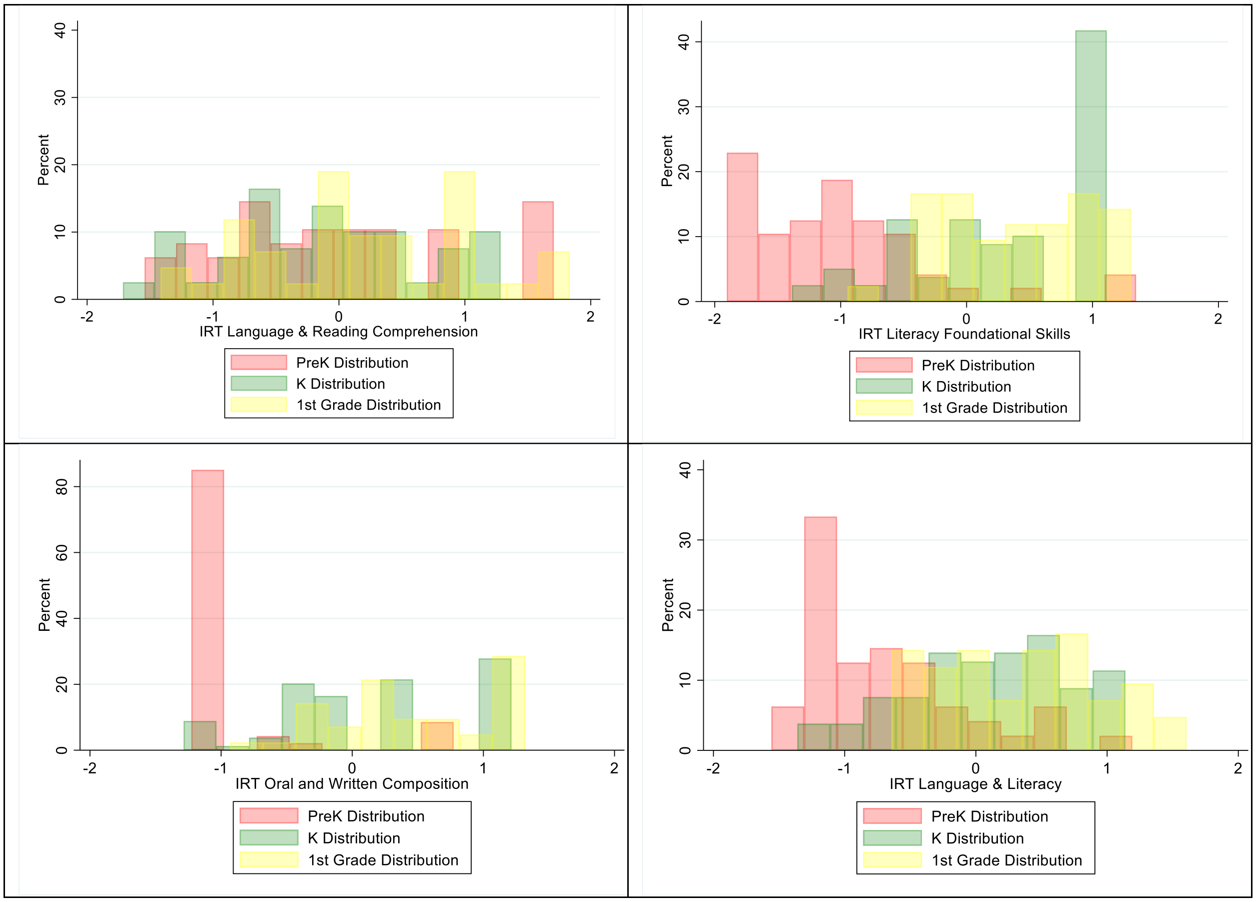


Figure 4D. Histograms of Language and Literacy Empirically-based Measures by Year

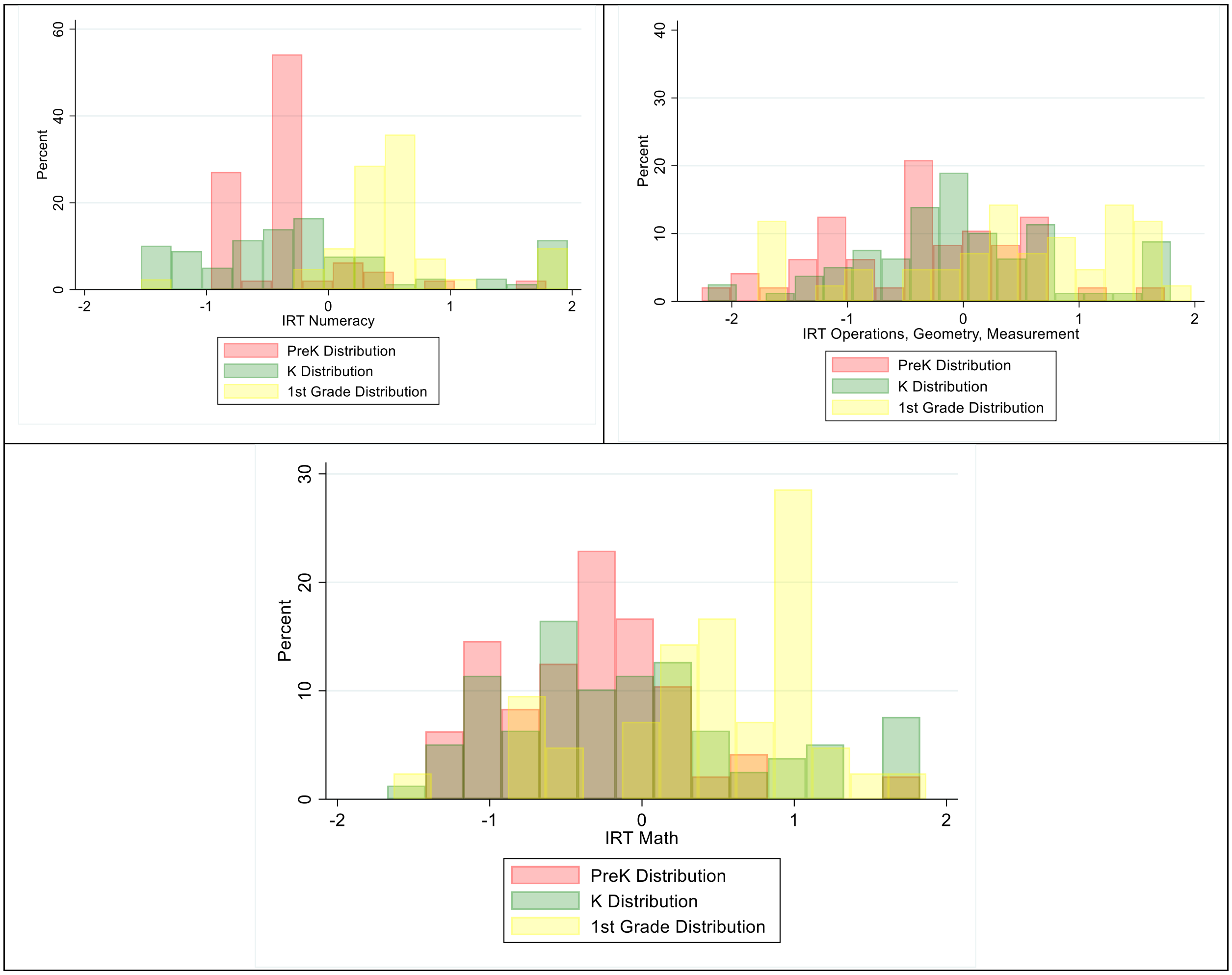


Figure 5D. Histograms of Empirically-based Math Measures by Year

Appendix E: Structural Quality Proxies

Massachusetts Quality Rating and Improvement System QRIS

The Massachusetts QRIS assess the quality of Family Child Care and Group / School Age programs, using standards customized for each type of care in the following categories: curriculum and learning; safe and healthy indoor and outdoor environments; workforce development and professional qualifications; family and community engagement; and leadership, administration, and management (*Learn about the Massachusetts Quality Rating and Improvement System (QRIS) - Mass.Gov*, n.d.). The same standards are used to measure school-based and center-based preschool programs.

To be rated at Level 1, licensed programs apply through the Massachusetts EEC QRIS portal that includes a self-assessment report based on licensing standards. To be rated at Level 2, programs submit supporting documentation for each of the rating factors presented above. Supporting documentation includes formal records of staff and administrators qualifications, experience, and professional development (PQ); internal documentation such as Internal Professional Development Plans (IPPDP), program plans for family involvement and transitions, and business plans; and self-administered scales such as the Program Administration Scale (PAS) and the corresponding ERS scale based on the center age-levels (ITERS-R for classrooms serving infants and toddlers and/or ECERS-R for classrooms serving three and four year old children). To be rated at Level 3, programs need to meet specific score thresholds in self-assessed instruments (PAS = 5 or higher; ITERS or ECERS = 4.5 or higher, with subscales scores of 3 or 4 depending on the subscale; a self-assessed CLASS score of 3 or higher in

Positive Climate, Reversed Negative Climate, and Teachers Sensitivity or Arnett Caregiver Interaction Scale self-assessed score of 3.0 or higher), receive a Level 3 Technical Assistance Visit, and provide program documentation on each corresponding standard. Centers can also provide proof of current NAEYC accreditation in lieu of documentation. To be rated at Level 4, programs need to provide documentation certifying that the program implements a Massachusetts approved curriculum (i.e., High Scope, Creative Curriculum, Opening the World to Learning, Resources for Early Learning), reliable rater scores for ITERS, ECERS, and the CLASS dimensions mentioned above, scoring six points or higher. Programs also need to provide results of Level 3 individualized Technical Assistance from a Program Quality Specialist site visit, and thorough documentation of the program administration, financial records, among others (Department of Early Education and Care, 2016).

Massachusetts Compliance with Licensing Standards

The Massachusetts Department of Early Education and Care licenses small group, large group, school age, and family care providers, family child care assistants, residential programs for children, and adoption/foster care placement agencies (*Child Care Program Licensing - Mass.Gov*, n.d.). We describe the process and measures for Large Group Providers only.

Programs that care for ten or more children on a regular basis outside a home need to apply for a Large Group license through the Early Education and Care regional office. To qualify for a license, providers need to meet health and safety, educational, and operational requirements. To begin with, potential providers take a mandatory training consisting of two online courses that cover licensing, the process to conduct background checks, pre-licensing, business considerations, and professional qualifications needed to apply for and operate a licensed Center-Based business. Potential providers obtain certificates of physical facility inspections including building, fire, water source, lead paint, and transportation if applicable. Potential providers also verify that hiring requirements, professional qualifications, and all staff protocols are met; and pay a fee ranging between $225 and $335 (depending on capacity) to obtain a provisional license. For licensed providers, renewal fees range between $275 and $450 (depending on capacity) and capacity increases fee is $75. Providers submit the required forms (e.g., child's enrollment, consent for child to leave the program, developmental history, staff schedule, etc.). Licensing policies and regulations are publicly available (https://www.mass.gov/lists/licensing-policies-for-group-and-school-age-child-care-programs) for programs review.

To begin the licensing application process, providers attend an in-person meeting and request credentials to access the EEC's Licensing Education Analytic Database (LEAD) system and place a licensing request submitting the required information. Once all required application materials are submitted, a licensor contacts the program representative and schedules a licensing visit. During the visit, the licensor verifies compliance with a set of indicators (*Mean* = 101, *SD* = 87, *range* = 22 – 275). Table 1B shows the maximum number of items assessed for each of ten licensing factors.

We estimated the percentage of compliance for each assessed factor, including only items with variability across centers to maximize discriminatory properties of our measure (see Table 2F). Due to small number of centers assessed on family involvement and transportation – only assessed if the center offers transportation services – we obtained the average compliance for centers across the remaining factors. We show the distribution for our measure of average compliance in Appendix F (Figure 7F).

Table 5E. Items Assessed and Items with Standard Deviation Higher than Zero

	Total Items Assessed	Items with Variability	% Total items with variation
Administration	55	33	60%
Interactions	25	12	48%
Curriculum	30	5	17%
Facilities	53	23	43%
Family Involvement	17	1	6%
Staff & Ratios	30	8	27%
Health and Safety	74	34	46%
Nutrition	19	13	68%
Transportation	23	9	39%
Total	326	138	42%

Table 6E. Descriptive Statistics of Compliance for Each Licensing Factor

Variable	Observations	Mean	Std. dev.	Min	Max
Administration	217	0.89	0.11	0.38	1.00
Interactions	215	0.99	0.07	0.00	1.00
Curriculum	197	1.00	0.04	0.50	1.00
Facilities	217	0.94	0.12	0.25	1.00
Family Involvement	36	1.00	0.00	1.00	1.00
Staff & Ratios	217	0.96	0.12	0.33	1.00
Health and Safety	217	0.86	0.16	0.29	1.00
Nutrition	203	0.98	0.07	0.50	1.00
Transportation	91	0.97	0.12	0.20	1.00
Total	217	0.94	0.05	0.69	1.00

Appendix F: Structural Quality Distributions and Geospatial Patterns

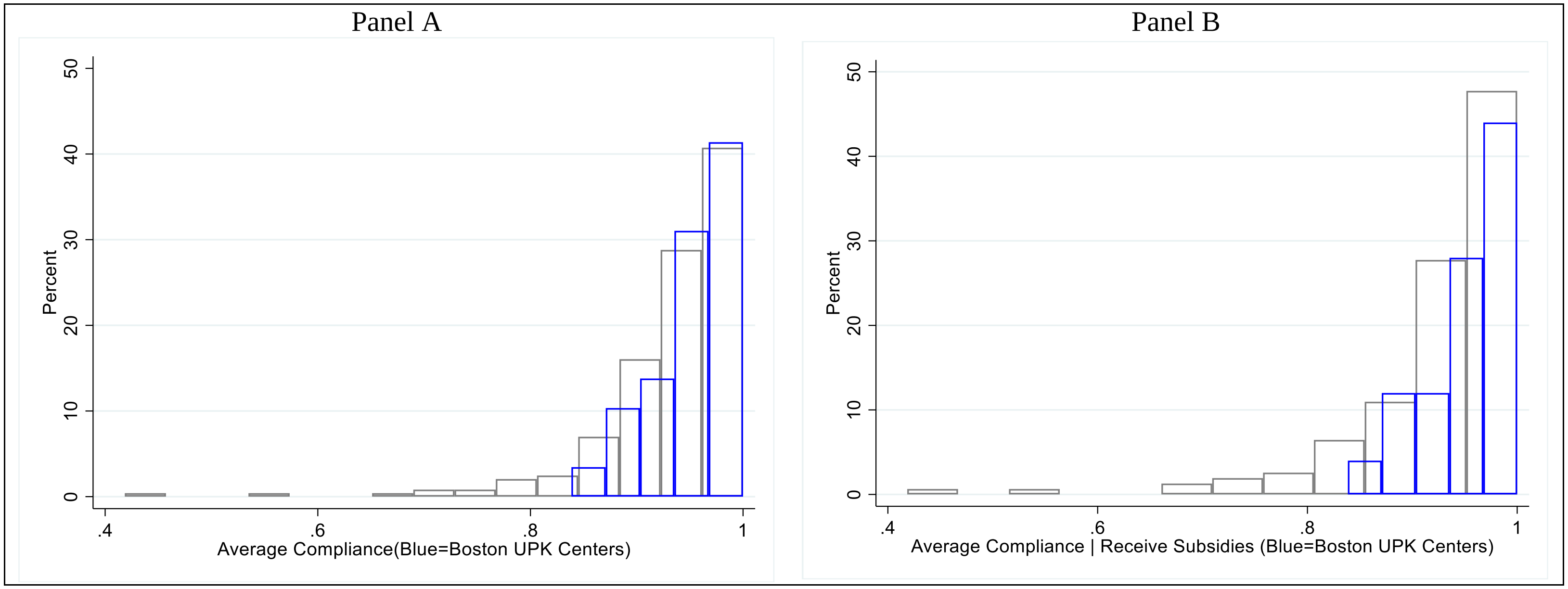


Figure 6F. Distribution of Compliance for all Centers (Panel A) and Centers Receiving Subsidies (Panel B)

Table 7F. Taxonomy of Linear Probability Models of Subsidy Receipt Status Among Boston Centers

	Center Receiving Subsidies in 2019			
	(1)	(2)	(3)	(4)
Center's capacity				
Capacity	0.00			-0.00
	(0.00)			(0.00)
Proxies of Structural Quality				
Center Participates in QRIS		0.68***		0.58***
		(0.05)		(0.07)
Average Compliance with Licensing Standards		-0.98**		-0.80*
		(0.37)		(0.38)
Community characteristics at the center location				
Children under 5YO			0.00**	0.00
			(0.00)	(0.00)
% Asian			0.01	0.00
			(0.00)	(0.00)
% Black or African American			-0.00	-0.00
			(0.00)	(0.00)
% Hispanic or Latino			0.00	0.00
			(0.00)	(0.00)
% Other and Mixed			0.00	0.00
			(0.01)	(0.00)
Estimate Median household income in the past 12 months			0.00	0.00
			(0.00)	(0.00)
% Speak other languages			-0.01*	-0.01
			(0.00)	(0.00)
% Bachelor's degree or higher			-0.01***	-0.01***
			(0.00)	(0.00)
Constant	0.52***	1.11**	1.20***	1.47***
	(0.10)	(0.35)	(0.20)	(0.33)
Observations	193	193	193	193
Neighborhoods	16	16	16	16

Note. Robust standard errors in parentheses. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$

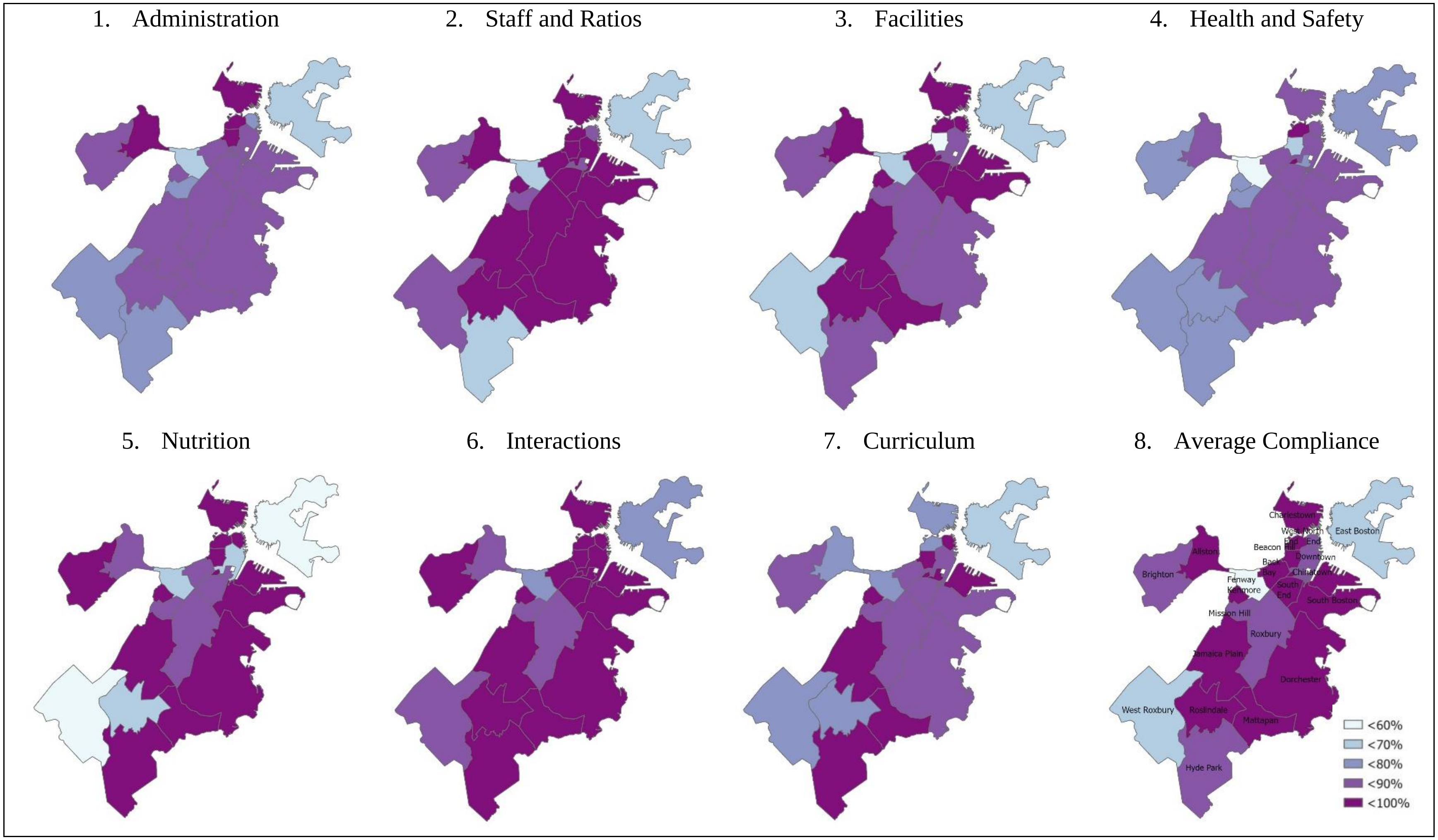
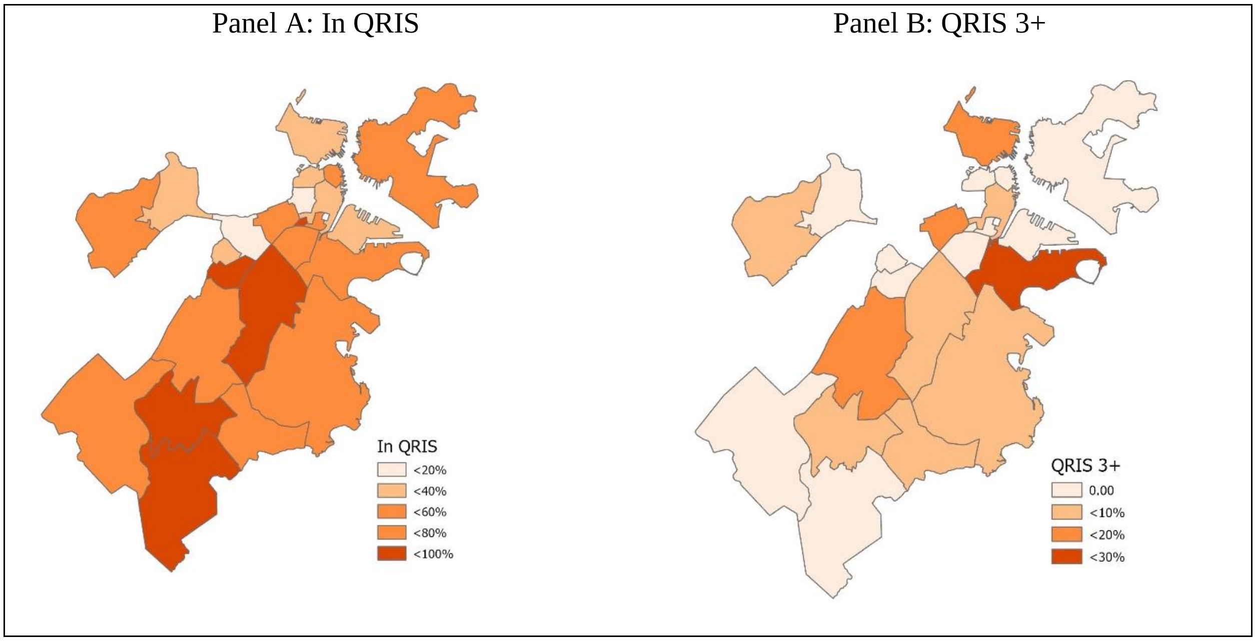


Figure 7F. Percentage of Compliance with Licensing Standards


Panel A: In QRIS

Panel B: QRIS 3+

Figure 8F. Proportion of Centers Participating in QRIS (Panel A) and Rated at Levels 3 or 4 (Panel B)