The Effects of Full-day Pre-kindergarten: Experimental Evidence of Impacts on Children's School Readiness

Allison Atteberry, Daphna Bassok, and Vivian C. Wong

Accepted for publication July 2019,

Educational Evaluation and Policy Analysis Journal

DAPHNA BASSOK is Associate Professor of Education and Public Policy at the University of Virginia, 405 Emmet Street South, PO Box 400277, Charlottesville, VA 22904; db9ec@virginia.edu. Her research focuses on early childhood education policy.

VIVIAN C WONG is Associate Professor of Research, Statistics, and Evaluation in the Curry School of Education at the University of Virginia, 405 Emmet Street South, PO Box 400277, Charlottesville, VA 22904; vcw2n@virginia.edu. Her research focuses on methodological issues related to causal inference and evaluating interventions in early childhood and K-12 systems.

ALLISON ATTEBERRY is Assistant Professor in Research and Evaluation in the School of Education at the University of Colorado-Boulder, 249 UCB, Boulder, CO 80309. allison.atteberry@colorado.edu. Her research focuses on policies and interventions that are intended to help provide effective teachers to the students who need them most.

Acknowledgments: This research was funded by Westminster Public School District. We thank them for their generous support. All errors are solely attributable to the authors.

1. Introduction

High quality early childhood education (ECE) programs can have a profound effect on children's development while simultaneously yielding substantial social returns (Blau & Currie, 2006; Heckman, 2006; Shonkoff & Phillips, 2000; Weiland & Yoshikawa, 2013; Wong, Cook, Barnett, & Jung, 2008). Further, the benefits of ECE are most pronounced for low-income children (Weiland & Yoshikawa, 2013), Latinx children (Gormley, 2008), and Black children (Bassok, 2010), suggesting that investments in ECE may be powerful tools for tackling early childhood achievement gaps and inequality. For these reasons, public investment in ECE has grown rapidly in the United States over the past two decades (Barnett et al., 2017).

Despite strong evidence that early investments *can* create large and lasting benefits for children, research tracking the impacts of large-scale, present-day ECE programs has yielded mixed results. Recent reviews of the literature indicate that children who attend public preschool in the year prior to kindergarten start school significantly ahead of their peers (Weiland, 2018). However, a growing body of research also suggests that the initial benefits on children's academic skills may be short-lived, dissipating quickly as children progress through school (Philips et al., 2017)

These findings have led to heightened interest among policy-makers and researchers in identifying specific program characteristics that promote returns on early childhood investments. In a recent consensus statement, a group of early childhood experts stressed the need to understand how preschool can serve as "an enduring base for future learning" and emphasized a need to unpack the particular features of preschool programs that contribute to children's development (Philips et al., 2017). Traditionally, policy-makers and researchers have focused on "structural" characteristics of ECE settings such as the qualifications of educators, the class size, and the staff-

child ratios. More recently, there also has been substantial interest in process-oriented features of ECE, such as quality teacher-child interactions, effective curricula, and access to professional development. Despite a growing literature on the role of these quality features, our understanding of the *causal* relationship between specific ECE features and child outcomes remains underdeveloped with little consensus on the features—or the combination of features—that are most critical for promoting children's development.

One salient characteristic—program intensity, or hours of exposure—has garnered considerable attention as a potentially important policy lever for supporting children's early learning. Between 1998 and 2010, the percentage of kindergarteners in the U.S. in full-day *kindergarten* grew rapidly from 55 to 80 percent (Bassok, Gibbs, & Latham, 2018). The percentage of preschoolers in full-day preschool also increased, albeit more slowly. In 2000, 47 percent of young children attended full-day programs. By 2016 that figure rose to 54 percent (Kena et al., 2016). Increasingly, policy-makers are exploring strategies to lengthen the school day in public preschool programs. For instance, the Office of Head Start proposed a new performance standard in 2015 that aimed to raise Head Start's operating hours from 448 hours a year to at least 1,020 hours per year (*Head Start Performance Standards*, 2016).

Efforts to increase children's hours of exposure in ECE settings have been motivated in part by the hypothesis that expanding the length of the school day will provide children with more exposure to high quality learning opportunities which, in turn, will yield greater and longer-lasting benefits. Full-day preschool programs might also attract new families who would otherwise not enroll their children in classroom-based ECE programs because their work or school schedules conflict with part-day programs.

Currently, there is little empirical evidence about the extent to which access to full-day versus half-day preschool yields large benefits, an important gap in the ECE literature given the relative cost of expanding the length of the preschool day. Full-day preschool expansion is expensive and has the potential to divert funds away from other ECE resources that may be more impactful in promoting children's development.

This study presents results from an RCT of full- versus half-day pre-kindergarten (pre-k) in a school system near Denver, Colorado. To our knowledge, this is the first rigorous randomized control trial (RCT) about the benefits of full-day, full-week preschool on children's school readiness skills. We find that, relative to an offer to attend half-day preschool, full-day preschool produces substantively meaningful, positive effects on children's receptive vocabulary skills (0.267 standard deviations) in the spring of their preschool year. Among those children enrolled in the public preschool program, full-day preschool also yields positive effects on teacher-reported measures of children's cognition, literacy, math, and physical development. Finally, our findings suggest that positive impacts are still evident as children start their kindergarten year. Combined, these short-term effects suggest full-day preschool programs had a meaningful impact on children's school readiness skills, suggest the *promise* for longer-term impacts.

Section 2 provides background about the potential benefits of intensifying children's exposure to ECE, Section 3 describes the context for the current study. Section 4 describes the study design, measures, and analysis models; Section 5 presents the impacts of full-day pre-k on children's outcomes at the end of pre-k as well as the beginning of kindergarten. We conclude by offering recommendations for policy-makers, and areas for future work.

2. Background

ECE programs vary substantially with respect to structural features (e.g. teacher education levels, ratios), process features (e.g. the quality of teacher-child interactions) and importantly, their contributions to children's learning (Bassok, Fitzpatrick, Greenberg, & Loeb, 2016; Morris et al., 2018; Weiland, 2018). Improving ECE at scale requires a better understanding of which particular features are most important for program effectiveness. Towards this goal, a growing body of research has examined the effect of specific program characteristics. For instance, recent RCTs have examined the effects of professional development for ECE teachers, as well as the impacts of specific curricula and teacher-child interactions (Araujo, Carneiro, Cruz-Aguayo, & Schady, 2016; Clements & Sarama, 2008; Early, Maxwell, Ponder, & Pan, 2017; Piasta et al., 2017). However, relatively few studies in ECE have used experimental methods to examine how structural features of ECE programs, which are the primary drivers of programs' costs, impact children's development. For example, there are no experimental studies measuring the impact of teacher education levels, teacher pay, or teacher-child ratios on children's learning in ECE settings. Similarly, few studies have provided rigorous causal evidence on the link between children's learning and the *intensity* of an ECE program, defined broadly to encompass both the number of years children attend a program and the number of hours they are enrolled per week.

The Role of Intensity in Preschool Classrooms

The intensity of ECE experiences may impact child outcomes in a number of ways. Most directly, if ECE programs provide more engaging and stimulating environments for children than they would otherwise experience, additional time spent in those programs may foster greater

benefits. On the other hand, there may be diminishing returns to time spent in ECE settings, and *too much* time in those settings may actually have negative impacts.

In addition to the stimulation and learning opportunities that ECE programs may provide, these programs also play an important role caring for young children and ensuring their safety while their parents work or attend school. ECE programs with longer hours may better align with parents' work schedules and thus reduce the number of ECE settings and transitions a child experiences regularly. This heightened stability may be beneficial for young children, as navigating multiple ECE arrangements is linked to more behavioral problems and greater rates of communicable illnesses (Morrissey, 2009, 2013; Pilarz & Hill, 2014).

Beyond the direct impact of ECE experiences for young children's learning, publicly-funded ECE programs may also benefit families. A large body of research has documented, for instance, that reductions in the cost of child care impact maternal employment (Bauernschuster & Schlotter, 2015; Cascio, 2009; Herbst, 2017; Malik, 2018). In addition, several studies show that the parents of children enrolled in Head Start, the largest federally-funded preschool program, engage more with their children (Bauer & Schanzenbach, 2016; Gelber & Isen, 2013) and attain higher levels of educational attainment (Sabol & Chase-Lansdale, 2015). In turn, these changes in parental employment, education, and parenting practices may benefit young children's development.

Although existing research has focused on the impacts of ECE access and participation broadly defined, access to more intensive ECE programs could, theoretically, be particularly beneficial for families, especially if these more intensive programs are free or low-cost. By providing greater child care coverage, full-day, publicly-funded ECE programs may save families

money, allow families to secure more stable employment with higher wages, and/or reduce stress. All of these changes are hypothesized to lead to benefits for young children.

Existing Evidence on Preschool Intensity

Despite the strong theoretical case for investing in more intensive ECE programs, the empirical evidence is limited. The most effective and rigorously evaluated ECE programs provided intensive interventions for children and their families. For instance, the Carolina Abecedarian Project, one of the most-touted ECE programs for its sizable impacts into adulthood, offered full-day preschool, five days a week, from infancy to age five and has been linked to positive outcomes through age 30 (Campbell et al., 2012; Campbell & Ramey, 1994). However, it is not clear whether these findings were *caused* by the relatively intense exposure or whether they could be explained by other features of the Abecedarian program. For example, from infancy, Abecedarian children were exposed to rich learning environments with trained child development specialists and health and medical professionals, while children in the Abecedarian control condition stayed home, without access to similar care environments. The existing literature fails to isolate the *unique contribution* of intensive exposure.

Unfortunately, there is only a small body of literature examining the impact of preschool intensity on children's development. Only one existing study is experimental, and it is unpublished (Robin, Frede, & Barnett, 2006). That study included 294 four-year-old children drawn from an urban school district serving mostly low-income families who were randomly assigned to full (N = 77) or half day (N = 217) classes. The half-day program consisted of 2.5- to 3-hour classes for 41 weeks; the full-day program consisted of 8-hour classes for 45 weeks. At kindergarten, full-day program children scored significantly higher on cognitive assessments compared to those in the half-day program and continued to outperform the comparison group at first-grade. However, the

full-day preschool group was more advantaged at baseline compared to their half-day counterparts, limiting the interpretability of the RCT results. For example, full-day children scored significantly higher on multiple pre-intervention assessments and their mothers worked more hours per week. Given the lack of baseline equivalence, the results from this study should be interpreted with caution.

All other studies exploring the link between preschool intensity and child outcomes rely on non-experimental methods and may not fully account for the non-random sorting of children into more intensive ECE programs (Herry, Maltais, & Thompson, 2007). For instance, Reynolds et al. (2014) compared outcomes of children who attended full- and half-day programs within the same school, and they found that children in full-day have better attendance and scored higher on four of the six school readiness indicators, including language, math, social-emotional development and physical health. However, the authors caution that their results may be biased because the full-day program prioritized enrollment for four-year-olds, so children in the half- and full-day programs were not equivalent with respect to age at baseline. Similarly, Gormley Jr, Gayer, Phillips, and Dawson (2005) show that Latinx children enrolled in full-day pre-kindergarten in Tulsa benefit more than those enrolled in half-day programs. However, they cannot disentangle whether this is because the full-day program is more effective or because of the non-random sorting of certain children and families into that program.

Overall, the findings from these correlational studies are mixed. While some studies find that the association between ECE participation and child outcomes is more pronounced for children who spend more hours in preschool per week (Loeb, Bridges, Bassok, Fuller, & Rumberger, 2007), others indicate that children who spend more hours in center-based child care

exhibited somewhat higher incidences of behavioral problems (Belsky, 2002; Vandell, Belsky, Burchinal, Steinberg, & Vandergrift, 2010).

Findings from the Federal Head Start program also yield mixed results. Using propensity scores and 2016 Family and Child Experiences Survey (FACES) data, Leow and Wen (2017) find no benefits of full-day classes on five academic and social outcomes in kindergarten. In contrast, in his reanalysis of data from the Head Start Impact Study, Walters (2015) found that Head Start centers offering full-day services produced larger impacts on children's cognitive outcomes compared to centers providing only part-day programming. Notably, however, this result may have been due to the offer of full-day services or other unobserved program feature related to full-day programming.

Finally, in a related set of studies, researchers have examined ECE intensity by comparing the benefits of participating in one versus two years of preschool. Using observational approaches, such as propensity score matching, researchers found that, relative to children with one year of preschool, those with two years showed improved performance both at school entry, and as they progressed through the early elementary grades (Leow & Wen, 2017; Shah et al., 2017; Wen, Leow, Hahs-Vaughn, Korfmacher, & Marcus, 2012).

Taken together, these set of studies provide mixed evidence about the impact of more intensive exposure to ECE programs, and they are limited by concerns about non-random selection into more intensive preschool programs.

Lessons from the K-12 Context

Although the research base on the impacts of *ECE* intensity is under-developed, related research from the K-12 context does provide support for the hypothesis that more intensive preschool programs may benefit children. For instance, a number of quasi-experiments indicate

that lengthening the school day leads to increases in children's academic outcomes (Battistin & Meroni, 2016; Bellei, 2009; Figlio, Holden, & Ozek, 2018). There is also a large body of research comparing outcomes for children enrolled in full- versus half-day kindergarten. The rapid expansion of full-day kindergarten in recent years has fostered heightened interest among policymakers and researchers in understanding how children are impacted by the longer school day. Given the age proximity of kindergartners and preschoolers, this line of research may be particularly relevant.

Unfortunately, here too the causal evidence is limited. Only one study uses random assignment to identify the impact of offering full versus half day kindergarten on children's outcomes. Gibbs (2014b) studied full-day kindergarten programs in Indiana, where lotteries were used to allocate over-subscribed full-day slots. Comparing children within the same school, she found that children randomly-assigned to full- rather than half-day kindergarten scored 0.31 standard deviations higher on a literacy assessment by the end of the kindergarten year.

To date, nearly all other studies tackling this question have relied on observational data, comparing children who attended full-day programs to those who attended half-day programs after accounting, to the extent possible, for selection factors at the child, family, school, or community level (Brownell et al., 2015; Gullo, 2000; Zvoch, Reynolds, & Parker, 2008). In general, these studies suggest positive but fleeting associations between full-day kindergarten participation and child outcomes. A meta-analysis of 40 studies of full-day kindergarten released between 1979 and 2009, for example, indicated that at the end of kindergarten, children who attended full-day kindergarten scored about a quarter of standard deviation higher than similar children in half-day programs, but that as children progressed through the elementary school years, these differences between groups disappeared (Cooper, Allen, Patall, & Dent, 2010).

Taken together, the K-12 literature does provide suggestive evidence that longer school days positively impact children, at least in the short-term. It is not clear, however, whether results from the kindergarten context generalize to preschool, as ECE programs serve younger children with unique developmental needs. Classroom practices, routines, and curricula differ between ECE and kindergarten classrooms, and the teachers guiding children's learning oftentimes differ substantially across these contexts with respect to their education, training, and compensation (Abry, Latham, Bassok, & LoCasale-Crouch, 2015; Whitebook, Phillips, & Howes, 2014).

Current Study

The goal of the current study is to provide rigorous evidence about the effects of one important and manipulable aspect of children's ECE experiences—program intensity. More intensive preschool programs are hypothesized to benefit young children both directly through increased exposure to a stimulating environment and indirectly through benefits for children's family. However, rigorous empirical evidence on these benefits is lacking, a major gap given the cost of funding expanded programs. The existing research base on full-day preschool is small and suffers from methodological limitations. Although there is a relatively larger literature on the closely-related question of full-day kindergarten, the causal evidence in this area is limited too, and findings from that context may not generalize to younger children. Our study adds to the existing literature by providing new experimental evidence about the impacts of full-day preschool on a host of short-term outcomes in a low-income, largely Latino population.

3. Study Context

Westminster Public Schools (WPS) is a public-school district located northwest of Denver that serves approximately ten thousand students annually. The district serves a population of students that is largely non-White (83 percent), low-income (76 percent), and non-native English

speaking (34 percent). While WPS is smaller than Denver, the percentage of students who are Latinx is larger (72 versus 59 percent in DPS) and the percent free/reduced price lunch (FRPL) eligible is about the same (68 versus 64 percent in Denver). In recent years, WPS has struggled to overcome the systemic socio-economic barriers that inhibit the performance of these students. While about 50 percent of WPS students perform at or above proficiency on statewide exams, there are large disparities in academic achievement between groups. For instance, while virtually 100 percent of WPS's fully English-fluent students are "proficient" in grade 3 math scores, only about 15 percent of its Not-English Proficient (NEP) students achieve proficiency status (WPS TCAP Results, 2014).

WPS leaders viewed ECE programs as one promising tool for addressing their students' needs. To intensify its ECE offerings, WPS used a pay-for-success funding model and secured funding to expand its pre-k program from half-day only to also include full-day classes among four-year-old children. Prior to the 2016-2017 school year, WPS provided only half-day preschool for three hours per day, four days per week. However, only about half of the district's eligible 1,100 four-year old children actually enrolled in the district pre-k program, leading district leaders to consider how to serve more Westminster families (Interview With Early Childhood Department Leadership, 2016). WPS hypothesized that many district families did not take advantage of WPS pre-k services due to the half-day and partial week program availability, which may have conflicted with family's child care needs. In the summer of 2016, WPS launched the Full-Day Pre-K Program ("FDPK") for the 2016-2017 school year. Because the district anticipated oversubscription[1] in its full-day classes, it held lotteries to award families with FDPK slots.

[1] In July 2016 alone, more than twice as many families applied for FDPK than could be accommodated.

Families who did not receive slots in the full-day program were offered enrollment in the business-as-usual half-day program.

To assess the efficacy of FDPK on student and family outcomes, WPS committed to a rigorous evaluation of its initiative. WPS worked with the research team to randomly assign offers of full- and half-day pre-k to eligible families. In 2016-17, the district opened seven new full-day pre-k classrooms that were available for six hours per day, five days a week; the half-day program ran as it had in past, with classes available for three hours per day, four days a week. Compared to the half-day program, FDPK more than doubled the number of hours per week for children in ECE settings and added more than 600 classroom-hours over the school year. In part, these additional hours were used for lunch and a daily nap. Beyond this, teachers could use the remaining hours in a variety of ways including literacy instruction, math instruction, structured or unstructured play, etc. Aside from the substantial differences with respect to intensity, full- and half-day classrooms were similar in many respects. All WPS pre-k classrooms were led by teachers with a bachelor's degree, they maintained the same teacher-to-child ratios, and they used the same curriculum , Little Treasures (which has now been replaced with "World of Wonders" by publisher McGraw-Hill). Little Treasures is a prekindergarten curriculum available online for free that has been used in WPS since before the study began. Little Treasures is described by its publisher as "a comprehensive, research-based Pre-K program" (Macmillan/McGraw-Hill, 2019). We are unaware of external evaluations of Little Treasures' efficacy, however a quasi-experimental evaluation of its counterpart, World of Wonders, did not find significant effects on student academic achievement in fourth grade reading (Corcoran, Eisinger, Kim, & Ross, 2016). Because it is free, Little Treasures is used widely across the U.S. Having more time to implement a widely-

available and widely-used curriculum with limited efficacy evidence is important for the interpretation of study results.

This paper presents findings from the first year of FDPK and is focused on children's school readiness outcomes, as measured both at the end of the preschool year and at the beginning of kindergarten. Below we describe the research design for the evaluation study, analysis models, and sensitivity checks for RCT estimates.

4. Methods

Research Design

The current study of FDPK employs a randomized block design (within first choice of school site) in which eligible families who completed an application were randomly assigned to offers of full- and half-day classrooms. In this case, the block randomized design was ideal because it allowed the study team to accommodate families' preferences for school sites, while also reducing the likelihood for chance imbalances across groups. Because some families did not take-up their lottery assignments into full- or half-day classrooms, we estimate the intent-to-treat (ITT) and complier-average-treatment-effects (CATE) as the causal estimands of interest.

Sample

All children who reside within the district and are age four by the first of October were eligible to participate in the FDPK program. In order for families to enroll in WPS generally, they needed to complete a required preschool application that includes health certifications and the child birth certificate. Families were included in the current study if they expressed interest in full-day preschool on their application, they completed the consent process, and their child had no known special education needs that prevented them from being served within a full-day classroom (e.g., if special equipment was required, a six-hour day is inappropriate).

Table 1 provides descriptive information about the study sample of 226 children (114 offered full-day, 112 offered half-day). Overall, the sample is largely Latinx (74 percent) and low-income (61 percent qualified for free-lunch, and 13 percent qualified for reduced-price lunch). The General Preschool Application (more on this data source below) asks the child's primary caregiver a series of questions about the child's family history. For instance, 37 percent reported having received some education beyond high school, and 49 percent indicated that their home language is not English. About 17 percent responded 'yes' to the question, "Has an immediate family member [of the child] received Special Education services?". About 23 percent of caregivers also indicated that the enrolled student has low language development, and 37 percent indicated low social development for the child.[2] The average age of children enrolled in the pilot study was 4.4 years, and about half the children are male.

Treatment Contrast

While many aspects of the full- versus half-day conditions were the same (e.g., same teacher training requirements, same curriculum, same professional development, same student-teacher ratio), students in these settings experienced a very different school year. Naturally, because full-day classrooms had 18 more hours of class time each week than did the half-day classrooms, the primary difference between the assigned treatment conditions was time allocation. In Table 2 we present descriptive statistics from the teacher survey on how teachers reported spending their time each day of a "typical school week". The largest differences between full- and half-day classroom time use are in the areas of napping (69 versus 0 minutes per day) and eating

[2] The exact wording of the language development question was: "Is your child in need of language development including, but not limited to, the ability to speak English?". The exact wording of the question about social development was: "Does your child have problems with social situations?". See footnote of Table 3 for the wording of all reported questions for parents on the general application.

(52 versus 17 minutes per day). Full-day students have a scheduled nap every afternoon, whereas half-day students do not (morning or afternoon sessions). Half-day students are also not served lunch: The morning session is from 8:00am to 11:00am, the afternoon session begins after lunch from 12:00n to 3:00pm, and the full-day session runs 8:00am to 3:00pm and therefore is the only setting that includes the 11:00- 12:00pm lunch period.

Turning to instruction, both classroom types allocate somewhat similar *proportions* of the class day to "academic" activities such as reading/literacy, math, social studies and science, but because full-day students are in class so much longer per week, those percentages add up to very different total number of hours exposed to these activities each week: Full-day students receive 3.7 hours per week of reading instruction (relative to 1.3 hours for half-day), 2.4 hours per week in mathematics (1.0 for half-day), and 1.4 hours of social studies and math (0.9 hours for half-day). Students in full-day classrooms also receive double the hours per week in non-academic activities such as visual/performing arts, play (structured and unstructured), and transitions between activities.

With respect to other differences in treatment versus control conditions, it is also worth noting that teachers were not randomly assigned to full- versus half-day classrooms so that the study design would more closely mirror realistic district staffing practices. When funding was secured for the full-day classrooms, positions were made open to both existing ECE teachers and new hires, alike. The pay for half- and full-day teaching positions are the same (half-day teachers cover both an AM and PM session each day). In conversations with district leaders, we know there was no systematic sorting of stronger teachers to full-day positions. In fact, teacher survey data documents that *half*-day teachers tended to have somewhat *more* teaching experience overall (16.8

versus 12.7 years), years of pre-k experience (12.8 versus 9.2 years), and years at the current school (7.0 versus 2.9 years).

Data Collection

To examine the impact of FDPK on children's outcomes, the study team assessed children's receptive vocabulary skills and administered an intensive developmental screener that identifies children who may need special education services. These assessments were conducted within the first month of fall 2016 (baseline) and again in the last month of spring 2017 (end of pre-k year). Crucially, all study children were administered the same assessments, regardless of whether or not they enrolled in WPS pre-k. If a child was enrolled in WPS pre-k (half- and full-day programs), the study team administered the receptive vocabulary assessment and developmental screener during regular pre-k hours. If a child was not enrolled in WPS, the study team met with the family directly to administer measures (either at a school or library).

In addition to outcome measures collected by the study team, the evaluation also includes assessments administered directly by the school district. However, because these measures are available only for the subset of study children enrolled in WPS (80 percent in pre-k), we treat these study results as exploratory.

Measures of Student Skills

Primary outcomes. Children's receptive vocabulary was measured by the Peabody Picture Vocabulary Test, 4th Edition (PPVT-4) (L. M. Dunn & Dunn, 2007). The PPVT–IV is a 228-item test in standard English administered by having children point to one of four pictures that best corresponds to a spoken word. The PPVT–4 Scale is norm-referenced and is widely used as a measure of children and adult's receptive, or heard, vocabulary. The PPVT-IV has strong psychometric properties with evidence for high reliability and validity (L. Dunn & Dunn, 2013).

Though the home language is not English for all study participants, all children were given the opportunity to attempt the PPVT (in English). Every child was administered a training set, placed into an age-determined basal set, and then given a raw score based on the ceiling item and total errors, which was transformed into a standardized score based on raw score and month, as prescribed by the assessment.

The Early Screening Inventory-Revised (ESI-R) is a one-on-one, 20-minute developmental screening tool that is appropriate for children from 3 years 5 months to 5 years 11 months (Meisels, Marsden, Wiske, & Henderson, 1997). The ESI-R is designed to identify the possibility of a learning condition that could potentially affect students' future school success. The measure evaluates children's developmental abilities in three domains of school readiness. To assess cognition and language, the child is given four tasks that allow her to demonstrate ability to comprehend language, express ideas, and reason and count. To assess visual-motor/adaptive reasoning, the child is asked to replicate patterns with blocks and copy with a drawing. To assess gross motor skills, the child is asked to jump, hop, and other physical coordination tasks. The three domain scores are then summed into a single raw score than can be used to identify children who may need to be referred for additional evaluation for special services. A Spanish language version of the screener is also available, and we administer the ESI-R in the child's primary language. Study children under the age of 4.5 were given the ESI-P (preschool) version, and children above that age were given the ESI-K (kindergarten).

We administered the ESI-R for a number of reasons. First, the district was particularly interested in whether full-day preschool could alter the need for costly special education services in early grades. Second, studies of the ESI-R indicate the instrument is both reliable and valid—a reliability for the ESI-P of 0.98 and 0.87 for the ESI-K (Meisels, Henderson, Liaw, Browning, &

Ten Have, 1993; Moodie et al., 2014). The ESI-R was normed on a sample of about 5,000 children across 60 classrooms in 10 states, including Head Start, public schools, private child care and preschools (Fantuzzo, Perry, & McDermott, 2004). The latter two of the three ESI-R domains described above were adapted and implemented in the widely-used Early Childhood Longitudinal Study-Kindergarten Class of 1998-99 (ECLS-K) dataset (Rock & Pollack, 2002). In addition, the ESI-R has been used in a number of other early childhood studies (Curenton, 2011; Fantuzzo et al., 2004; Luo, Jose, Huntsinger, & Pigott, 2007). Finally, as shown in Table 1, students in our sample exhibited variation in their ESI-R scores (mean of 19 and SD of 6.7 in fall, mean of 20.6 and SD of 5.1 in spring), suggesting the items capture heterogeneity in students' skills during this age period.

Exploratory outcomes. As discussed above, in addition to the data collected directly by our research team, we also considered two outcomes collected by WPS. First during the fall and spring of the pre-k year, all teachers assessed children using Teaching Strategies GOLD (TS GOLD) a widely-used, observation-based authentic assessment (Heroman et al., 2010). Teachers observe children's skills during typical classroom sessions and evaluate them across up to nine broad areas of development (e.g., literacy, mathematics, language, social-emotional, cognitive, and physical). TS GOLD has been used in other studies tracking the association between preschool intensity and child outcomes (Reynolds et al., 2014). Although teacher-reported, the measure has shown strong reliability and validity in developer-conducted studies (Teaching Strategies, 2011, 2013), and two recent studies provided evidence of concurrent validity with direct assessment of similar skills[3] (Miller-Bains, Russo, Williford, DeCoster, & Cottone, 2017; Russo, Williford,

[3] To explore concurrent validity, both Miller-Bains et al. (2017) and Russo et al. (2019) correlate TS GOLD domain scores with other direct assessments of the same constructs. Miller-Bains et al. report in their Table

Markowitz, Vitiello, & Bassok, 2019). Although questions remain about the measurement properties of GOLD, recent research indicates the assessment functions well with children whose home language is not English (Kim, Lambert, & Burts, 2013).

The second district-administered measure we use in this study is the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) in kindergarten (Good & Kaminski, 2002). The DIBELS assesses children's early literacy skills in the areas of phonemic awareness, phonics, reading comprehensive, fluency, and vocabulary. It is designed to help identify children who may experience difficulty acquiring basic early literacy skills. The measure shows adequate validity and reliability and has been adopted by a number of states to assess children's school readiness (Good et al., 2004). WPS administers the DIBELS to children during the fall and spring of the kindergarten year and provided us with two subdomain scores—First Sound Fluency and Letter Naming Fluency—as well as an overall composite score. The current study presents RCT results on the TS GOLD at the end of the pre-k year, and the DIBELS in the fall of the kindergarten year.

Baseline Measures

A unique strength of the current study is that we have access to an unusually rich set of baseline covariates, including measures of child, home, and family characteristics as well as baseline assessments for primary outcomes. All applicants to Westminster complete a General Preschool Application, which included questions about children's race/ethnicity, gender, birthdate and age, free/reduced-price lunch program eligibility, primary language, and the primary language spoken in the home. In addition, the child's parent or primary guardian indicated their educational background, and whether there is a history of family drug/alcohol abuse, special needs, frequent

8 a correlation of 0.68 for literacy (0.53 in Russo et al.'s Table 3), 0.56 for math (0.44 in Russo et al.), and 0.40 for language (0.37 in Russo et al.).

school moves, housing difficulty, domestic abuse, social services involvement, and extreme child medical events occurring within the family. Parents also indicated whether they were concerned about the child's low language and/or social development. Finally, our research team administered measures of the PPVT and the ESI-R in early fall of the pre-k year and teachers also assessed children using the TS GOLD during the same period. These baseline covariates greatly enhance our ability to assess covariate balance across groups and improve statistical precision to detect effects.

Analysis Model

In Equation (1), we present the statistical model used to estimate causal effects of full-day pre-k on student outcomes of interest:

$$Y_{ijt=1} = \beta_0 + \beta_1(T_{ij}) + (\boldsymbol{X_{ij(t=0)}})\boldsymbol{\beta} + \alpha_j + \varepsilon_{ijt}. \quad (1)$$

Y_{ijt+1} represents an outcome for student *i* in school site *j* at the end of the second semester of the pre-k year (t=1). The variable T_{ij} is the dummy variable coded to 1 if the child was randomly assigned to receive an offer of full-day pre-k (treatment) and 0 if offered a half-day spot (control). Our analysis models also include a series of first choice school site fixed effects α_j, as well as a vector of time-invariant ($\boldsymbol{X_{ij}}$) and time-varying controls ($\boldsymbol{X_{ijt=0}}$) that were taken at baseline (these include the first fourteen variables in Table 1). In this model, $\beta_1$captures the intent-to-treat effect of interest—that is, conditional on covariates in the model, β_1 is the average effect on Y due to randomly assigned offers of full- or half-day pre-k. We also use two-stage least square approaches to estimate the complier average treatment effects—that is, it is the effect of *attending* full-day pre-k on children's school readiness skills.

Validity Checks for Experimental Design

Because randomized experiments in field settings are rarely (if ever) implemented perfectly, we conducted a series of diagnostic probes to assess the extent to which validity threats occurred. Below, we discuss the results of our diagnostic checks for covariate balance, treatment non-compliance, and for missing data on outcomes.

Covariate balance. Even with random assignment, it is possible the full- and half-day groups differ based on chance alone. We tested whether there were differences in groups at baseline by fitting a series of regressions in which each baseline covariate was regressed on indicators for whether the family was offered full-day pre-k and site fixed effects. The dependent variables in these regressions included all baseline covariates discussed above. Each row of Table 3 presents the results from a separate regression with the same right-hand side specification but a different baseline covariate as the dependent variable (logistic regression was used for binary covariates). The results in Table 3 are presented both in standardized difference metrics (Cohen's D), as well as their original metrics (e.g., percentages, mean scores).

The full- and half-day groups were similar at baseline. None of the group differences on the fourteen covariate outcomes are statistically significant. Following Ho, Imai, King, and Stuart (2007), the What Works Clearinghouse (WWC) uses a threshold of 0.25 standard deviations in absolute value (based on the variation of that characteristic in the pooled sample) as an upper bound for non-equivalence (What Works Clearinghouse, 2014). Again, none of the differences in Table 3 are above that threshold.

The groups are very well matched with respect to racial composition, age, gender, home language, and parental education. Furthermore, there does not appear to be any systematic patterns of advantage or disadvantage between the groups. The full-day group has somewhat higher

percentages of some characteristics that are historically associated with lower test score performance (e.g., children in the full-day group are approximately 12 percentage points more likely to be eligible for free lunch). However, on other variables, half-day families exhibit slightly higher means (e.g., 75.0 percent of control families are Latinx, while 72.8 percent of treatment families are Latinx). Finally, we fit a multivariate regression model for the full set of covariates as simultaneous outcomes, and we find that the F-statistic associated with the null hypothesis that the coefficient on the treatment variable is the same across all outcomes is not statistically significant ($F(15, 204)= 1.28$, p-value= 0.215). Overall, the balance tests suggest little evidence of systematic differences between groups at baseline. Further, sensitivity analyses (presented below) indicate that the magnitude of our experimental estimates is robust to the inclusion of baseline covariates in the model.

Treatment non-compliance. Families are randomized to *offers* of full or half-day pre-k slots in WPS. Some may choose to not take up their offers. In particular, families assigned to a half-day WPS slot may be more likely to opt out of WPS pre-k and could possibly enroll their child in a different full day setting. Such treatment non-compliance leads to a discrepancy between assigned and observed treatment status. Across both conditions, 74 percent of the study sample participated in the pre-k classroom to which they were assigned. Among those who were randomly assigned to full-day pre-k, 86 percent attended the full-day program in WPS. Among those assigned to the half-day group, 62 percent participated in half-day classes in WPS. This differential take-up rate across groups was expected given that all study families had initially indicated interest in a full-day slot. A small portion of study participants experienced crossover: Specifically, 2 percent of families assigned to full-day pre-k switched to the half-day program in WPS, and 9

percent of families who were initially assigned to half-day pre-k enrolled in the WPS full-day program.

One might be concerned that differential uptake (compliance) could bias our findings. Indeed, if treatment non-compliers were less advantaged at baseline or control non-compliers were more advantaged at baseline, then our CATE estimates would be inflated. Although this assumption is not directly testable, we find that, overall, there are no significant differences in pre-treatment covariate means between those who do and do not take up their randomized offer (for all 14 pre-treatment covariates).[4] Specifically, among the control group, none of the 14 pre-treatment covariate means differ by uptake status. For the treatment group, only 1 of the 14 covariate mean differences is statistically significant (the mean standardized fall PPVT score is 0.235 SDs and -0.816 SDs for those who do and do not uptake their offers, respectively). Taken together, this suggests that--though control group families are less likely to take their offer (differential rate), the decision to uptake is not systematic in terms of baseline covariates (but we cannot rule out unobserved predictors of uptake).

To address threats to internal validity from differential uptake, we estimate and focus on ITT effects that isolate the causal impact of receiving an *offer* to participate in the full-day program, as well as CATE effects, which estimate the impact of the program for compliers.

Missing data and attrition. Because families were required to complete the General Preschool Application to enroll in FKPK, there is very little missing data on baseline covariates (see Table 1 for missing data rates). For the few cases in which we did not have baseline covariate information, we included controls for missing data in the analysis models.

[4] Analyses available upon request.

Table 1 shows that 6.2 percent of the study sample is missing a spring ESI-R score, and 11.5 percent are missing a PPVT score—our two main outcomes of interest—because we were unable to reach some families to complete the assessments or the child refused to complete the activity. Table 4 presents again the baseline characteristics of half- and full-day children for the full sample (left), alongside the same descriptives for the remaining sample that does not have missing outcome data on which we conduct our main analyses. It shows that the remaining sample is similar to the complete study sample, and that there is no evidence of differential attrition across the groups. When we re-apply a multivariate regression model to the non-attrited sample for the full set of covariates as simultaneous outcomes, we again find that the F-statistic associated with the null hypothesis that the coefficient on the treatment variable is the same across all outcomes is not statistically significant ($F(15, 190)= 1.25$, p-value= 0.238).

By design, there is more missing data on assessments that were administered by the school district rather than by our study team. WPS administers TS GOLD to children during the fall and spring of the pre-k year, and DIBELS during the fall of the kindergarten year. Because the assessments are given only to participants who enrolled in public pre-k and kindergarten, we lack these outcomes for children who did not participate in WPS pre-k and/or kindergarten. As shown in Table 1, we lack TS GOLD data for about 20 percent of the study sample, and DIBELS data for 38.5 percent of the sample. When we use these district-administered assessments as exploratory outcomes, missing data rates are higher than for assessments collected by the study team, which suggests more opportunity for baseline imbalance.

Tables A1 and A2 in the Appendix A compare baseline information for the full RCT sample, and the subset of children who had TS GOLD scores at the end of the pre-k year and DIBELS scores in the fall of the kindergarten year, respectively. Reassuringly, we find that

although there is substantial missing data across both groups, we do not observe systematic differences in either full- or half-day group. As was true for the full sample, the treatment versus control group mean differences in the non-attrited samples—expressed in Cohen's D in the final columns of Tables A1 and A2—are not statistically significant and are not larger than the 0.25 WWC threshold. Moreover, the F-statistics for their multivariate regressions (shown in the table footnotes) are not significant. This lends some support to the possibility to the idea that the kinds of participants missing data may generally not systematically different from the full sample. Nevertheless, given the high levels of missingness, we treat these outcomes as exploratory. In specification checks, described below, we also assess whether impacts on these outcomes are robust to conservative assumptions about the nature of the missing data.

5. Results

Primary Outcomes

Table 5 presents our estimates of the ITT and CATE for our primary outcomes, PPVT and ESI-R. Columns labeled M1 show the standardized mean differences in effects with only school site fixed effects included in the model; the M2 columns show impact estimates with school site fixed effects and controls for student and family demographic factors; and the M3 columns present results from our preferred model, which includes school site fixed effects, controls for demographic factors, as well as baseline pretest scores. Across all three models, treatment effect estimates are generally stable, while the proportion of variance explained increases across models.

Table 5 shows that the offer of a full-day pre-k slot resulted in an increase of 0.275 standard deviations on the PPVT-4 (upper-left panel of Table 5, Model 3). The impact of actually attending full-day pre-k improved children's PPVT scores by 0.363 standard deviations (lower-left panel of Table 5, Model 3). For ESI-R, all estimates across models and estimands are positive and between

0.101 and 0.185 standard deviations. The ITT effect is 0.101 standard deviations and the CATE result is 0.132 standard deviations (upper- and lower-right panels of Table 5, Model 3). Neither are statistically significant.

Exploratory Outcomes

Table 6 contains ITT and CATE results for TS GOLD and DIBELS, the outcome measures collected only for those children enrolled in WPS. For the sake of parsimony, we only present results in Table 6 from our preferred Model 3. Table 6 results suggest that children randomly assigned to an offer of full-day pre-k were rated more highly on the TS GOLD than their peers in half-day programs. Looking holistically across subdomains, the treatment effects on overall TS GOLD scores—calculated by taking the mean of the six standardized subdomains (Russo et al., 2019)—are 0.258 standard deviations (ITT) and 0.320 standard deviations (CATE). Treatment effects are positive for all six domains assessed, and statistically significant for five of the six domains (cognition, literacy, math, physical development, and socio-emotional development). The largest effects were for literacy (ITT = 0.393 standard deviations; CATE = 0.487 standard deviations), followed by cognition (ITT = 0.258 standard deviations, CATE = 0.320 standard deviations), physical development (ITT = 0.237 standard deviations, CATE = 0.294 standard deviations), and math (ITT = 0.230 standard deviations, CATE = 0.285 standard deviations). TS GOLD scores on language are substantively meaningful and positive but do not differ significantly between groups.

Table 6 also indicates that by the fall of kindergarten children randomly assigned to an offer of full day pre-k outperformed their peers on the DIBELS. The ITT effect for the Overall Composite score of the DIBELS was 0.344 standard deviations, and for the CATE, it was 0.392 standard deviations. We also see positive estimated effects on the two provided DIBELS

subdomains—First Sound Fluency (ITT effect= 0.266) and Letter Naming Fluency (ITT effect= 0.354). All DIBELS effects were positive and substantively meaningful, though only the Overall Composite score and Letter Naming Fluency effects were statistically significant at the 5 percent alpha level.

Robustness Checks for Missing Outcome Data

Despite the fact that we make every effort to assess all study children not enrolled in WPS, we do not observe outcomes for all study children at the end of preschool or the start of kindergarten due to the natural mobility that occurs in any district during and after preschool (see Table 1). One could be concerned that the missingness is systematic and could bias our results. We would be particularly concerned in a scenario where high-scoring control students (or low-performing treatment students) were more likely to have missing data, as these patterns would bias our estimates upward so that they appear larger than they actually are.

As a robustness check for our estimated effects, we make assumptions about the missing outcome scores that would work strongly *against* our findings: Within each school, we assume that every missing control group child would have performed on these assessments at the average level of the apparently higher-scoring treatment group. Likewise, we assume that every missing treatment group child would have performed at the average level of the control group. These strong assumptions correspond to the upwards bias scenario described above. Recall that we do see evidence that the observable characteristics of the pre- and post-missing data samples are systematically different from one another, so this thought experiment may be somewhat overly-punitive. Nevertheless, we can look to see if the direction and magnitude of estimated effects under these assumptions remain positive and substantively meaningful (if not still statistically significant).

When we make these assumptions, we find the pattern of the results generally persists. We reproduce the analyses presented in Table 5 and Table 6 now with the imputed outcome data and present updated results in Table 7. Note that in Table 7 there are now N=226 children in every model because all study participants now have a value for the outcomes, imputed or otherwise. In Table 7 we see that the estimated effects are generally smaller in magnitude, but all remain positive and most substantively meaningful. Statistical significance should be interpreted with caution when analyzing imputed outcome data, however fourteen of the estimated effects continue to be statistically significant. This suggests that the direction and magnitude of our findings are insensitive to the missing data that is both present in this study and endemic to all longitudinal early childhood research designs.

6. Discussion

To complement the large body of research examining *whether* preschool leads to benefits for children, evidence is needed on the *conditions* under which preschool is most effective. The current study provides the first rigorous evidence on the effects of full-day, full-week preschool on young children's school readiness. Unlike the majority of the existing research on the intensity of early childhood interventions which reports regression adjusted associations between program exposure and child outcomes, the current study leverages a school-based lottery to conduct an RCT, thus isolating the true impact of an offer for this full-day, full-week program on young children's early development.

The results indicate that the offer of full-day pre-k has a positive impact on young children's school readiness skills. In particular, children offered full-day pre-k scored a quarter of a standard deviation higher on the PPVT—a widely-used measure of receptive vocabulary—than peers offered half-day pre-k.

This effect is substantively meaningful. To put it in perspective, we compare the effect size to the magnitude of impacts from rigorous studies measuring the *overall* impact of ECE interventions, arguably a stronger contrast than that explored in the current study which examines the added-value of a more intensive preschool program. Evidence from the experimental Head Start Impact Study showed that the effect of random assignment to Head Start on the same outcome considered here, was 0.18 for 3-year-olds assigned to Head Start, and 0.09 for 4-year-olds (Puma et al., 2010). Wong et al. (2008) used regression discontinuity methods to estimate the impact of five state pre-kindergarten programs on the PPVT. They found that effects sizes ranged from a statistically insignificant -0.13 in Michigan to a statistically significant 0.36 in New Jersey. Only two of the five states considered showed statistically significant positive impacts on this outcome. In a recent study expanding this work to eight state pre-kindergarten programs (Barnett et al., 2018), the average effect sizes of pre-kindergarten on the PPVT was 0.24, though only three of eight states (New Jersey, Michigan and Oklahoma) showed statistically significant impacts in the authors' preferred model, and results were sensitive to model fit. Finally, findings from the recently published evaluation of Tennessee's pre-k program indicate that at the end of preschool the ITT effect on a composite measure including language was 0.24, with a TOT effect size of 0.395 (Lipsey, Farran, & Durkin, 2018). Lipsey et al. specifically called out the 0.25 threshold as "educationally meaningful, e.g., by the 0.25 threshold used by the U.S. Department of Education What Works Clearinghouse" (pg. 165).

The effects observed in the current study are thus larger than what is often observed in studies measuring the *overall* impacts of ECE programs, and roughly the same size as those seen for some of the most successful state pre-kindergarten programs (Lipsey et al., 2018). These findings are encouraging, especially given the importance of unconstrained skills, and particularly

early vocabulary, for children's reading at third grade and longer-term literacy success (Snow & Matthews, 2016). As discussed in greater detail below, there are important differences in the populations of interest between the current study (largely Latinx) and the preceding ECE impacts studies that may be related to differences in the magnitude of effects.

Although we consider our findings on the PPVT to be our primary results, we do also find suggestive, positive results from the other outcome measures considered. For instance, we find positive but statistically insignificant effects at the end of the preschool year on the ESI-R, a developmental screener used to identify children who may need special education services. We also find positive outcomes with respect to the TS GOLD, a widely-used teacher-reported observational tool, which captures development across a broader range of developmental domains. Here too effects were encouraging. In particular, we find statistically significant and sizable impacts on five of the six subdomains and the overall score, ranging in ITT effects sizes from 0.15 to 0.39. The ITT coefficient for the language subdomain was about 0.11 but was not statistically significant.[5] A recent non-randomized study exploring the impacts of full-day preschool in the context of Chicago's Child-Parent Centers also showed that children in full-day classrooms outperformed their peers on four of six TS GOLD outcomes, though they found statistically significant outcomes on language and socio-emotional development but not literacy or cognition (Reynolds et al., 2014).

[5] It is curious that the PPVT effect is large and significant (0.275 SDs), while the GOLD subdomain with the smallest estimated effect is language (0.114 SDs). To rule out the possibility that this discrepancy can be explained by differences in the analytic samples (PPVT is given to all study children who can be reached (N=200), while GOLD is only administered to the subset of about 180 study children enrolled in WPS), we re-estimate the PPVT effects on the subsample of student with TS GOLD scores. However, we find that the estimated ITT effect on PPVT is 0.299 SDs in the GOLD sample, which is not very different and not smaller than the original estimate (results available upon request). Therefore, compositional differences in PPVT versus the GOLD sample do not explain why the language/ receptive vocabulary effects on the PPVT are larger than those captured by GOLD language

We interpret the TS GOLD findings with caution for two reasons. First, because the measure was collected as part of "business-as-usual" practice for WPS pre-k, it is unavailable for the non-random sample of children who ultimately did not attend pre-k in WPS despite receiving an offer to do so (about 20 percent of the study). Encouragingly, the findings in the current analysis are robust to relatively conservative assumptions about the values of this missing data. A second concern about the TS GOLD data is that it is reported by teachers. It is difficult to know how problematic this is. On the one hand, existing research suggests some concurrent validity of these measures to direct assessments of children's skills (Miller-Bains et al., 2017; Russo et al., 2019). WPS teachers have been routinely administering TS GOLD in their classrooms for at least six years prior to the study and therefore it is standard practice and not specific to this study. On the other hand, teachers are aware of children's full versus half-day status, and this knowledge may introduce bias. Relatedly, teachers who spend more time with children in full-day classrooms, may have more opportunities to observe children's skills relative to those in half-day classrooms. This too may introduce bias. The formal TS GOLD trainings teachers take in WPS are designed specifically to increase accuracy and reduce rating bias. Still, we cannot directly assess the existence or size of this bias in teacher-reported assessments, and therefore treat these findings as suggestive.

Finally, effect sizes for the DIBELS, a direct literacy assessment administered by WPS in the fall of the kindergarten year, are substantial (ES=0.34), though only marginally statistically significant given the smaller sample size. These results closely align with Gibbs (2014a) whose lottery-based analysis of full-day kindergarten shows ITT effects of approximately a third of a standard deviation. Again, these findings are encouraging given the association between early literacy, as measured by this assessment, and children's development of reading skills throughout

elementary school (Burke, Hagan-Burke, Kwok, & Parker, 2009; Rouse & Fantuzzo, 2006). However, here too, caution is warranted, given the relatively high rates of missingness on this WPS-only outcome.

Taken together, the effects documented in the current paper, which were systematically positive, and in most cases also statistically significant, provide the most rigorous evidence to date on the impacts of an extended pre-kindergarten day for young children's school readiness skills. These findings are important, especially in light of recent calls for more rigorous evidence on the impacts of specific aspects of ECE in fostering children's learning gains (Weiland, 2018). Before turning to the policy implications of the current results, we first highlight some important study limitations, as well as key questions our study *cannot* answer.

Limitations

Several aspects of the current analysis pose important limitations. The first is that only two of the outcomes considered in the study (the PPVT and ESI-R) were directly-assessed by the research team as part of the study and administered to all children irrespective of whether or not they enrolled in the study district. As discussed above, the TS GOLD and the DIBELS assessments are collected as part of "business-as-usual" practices in the district and are therefore limited to the 80 percent of study children who enrolled in WPS (for preschool). Although we conduct analyses to evaluate the sensitivity of our results to the non-random sorting of children into WPS across the treatment and control group, the study would benefit from a broader array of researcher-collected measures, or measures collected for all children. In particular, in light of research both about the effects of ECE programs on young children's social skills, and about the potential impacts of long days in child care on children's behavior, this study would benefit from more reliable measures of children's behavior and non-cognitive outcomes.

A closely related concern is the unsurprising, differential take-up of WPS preschool across the treatment and control group, and related issue of missing data. In the current study 86 percent of children offered a full-day slot enrolled in WPS compared to 62 percent of those offered a half-day slot. Although we have carefully considered the implications of this non-random sorting on our findings, we cannot fully account for bias that may be introduced into our analysis here.

Finally, a third potential limitation in our current study is the non-random sorting of teachers across half and full-day classrooms. Randomizing teachers to half- or full-day programs was, understandably, viewed as impractical by our district partners. This leaves the possibility that teachers assigned to teach in the more intensive classrooms differed in important ways from those assigned to half day programs, and that those differences rather than the intensity itself, is what is driving the impacts we document in the current study. Although our examination of observable teacher characteristics suggests clear sorting is not present and, if anything, half-day teachers may be more experienced, unobserved differences may still be at play.

Questions the Study Currently Cannot Answer

Beyond these data limitations, the current study, which focuses on the immediate impacts of full-day pre-k on child outcomes within one Colorado district, leaves many important questions unanswered. Four of these questions warrant particular consideration.

First: To what extent will the benefits observed at the end of the pre-k year and at the beginning of kindergarten be maintained as children proceed through the early grades and beyond? In recent years, concerns about the rapid "fade-out" of early childhood program effects has been a major concern among early childhood researchers and policy-makers. Often, the benefits observed from ECE programs at school entry, dissipate quickly as children progress through elementary school (Bassok et al., 2018). To get at the persistence of the effects documented in the current

study, we are tracking the children in the current study as they proceed through at least the first four years of elementary school (and up to six years). In addition to the current cohort of children, we are tracking two additional cohorts of WPS children, who will also be randomly assigned to an offer of full- or half-day pre-k. By following these children through a minimum of third grade, we will be able to track whether the initial benefits fade-out as children proceed through school.

Second: To what extent do the experimental findings documented in the current study—which focused on a predominantly Latinx, predominantly low-income sample—generalize to other contexts? Given the central role of replication in the accumulation of scientific knowledge, it essential to assess whether the findings documented in the current study replicate in other contexts (Duncan, Engel, Claessens, & Dowsett, 2014). Because there are so few non-Latinx children in the current sample (N=26), we are unable to effectively compare estimated effects between White and Latinx students. Districts serving predominantly Latinx children may benefit differentially from full-day preschool programs. Existing research suggests that, in general, the benefits of preschool participation are greatest among Latinx children. The same may be true for full-day preschool relative to half day. It may be, for example, that for English language learners, a primary way in which more hours in preschool lead to better outcomes is by providing more hours of English-language exposure. If this is a key mechanism, results may look different in communities with fewer English language learners, and future replications should measure the impact of full-day preschool in districts serving other populations of children. Such studies would inform whether policy-makers might prioritize targeted or universal full-day expansions, and they may also inform the extent to which moving towards full-day programs may ameliorate achievement gaps.

Third: If full-day classrooms are effective in supporting young children's learning, what specific practices and experiences are driving the benefits? Just as it is important to unpack the

mechanisms that lead to benefits from ECE program participation broadly, it is essential to understand the specific pathways through which full-day programs lead to benefits. Two broad categories of mechanisms may be at play. First, it may be that full-day programs offer children more stimulating learning environments than they would otherwise experience. Second, benefits to children may operate through effects on families, such as increases in work hours and earnings or decreases in stress. For the second and third cohorts of this study, we are collecting data that will allow us to explore these possibilities. In particular, through multiple classroom observations throughout the year, we are collecting detailed information about the time-use in full- and half-day classrooms, as well as the quality of teacher-child interactions as measured by the Classroom Assessment Scoring System, CLASS (Pianta, La Paro, & Hamre, 2008). We also supplement these observational measures with detailed parent surveys which allow us to explore the potential impacts of full-day programs on parental employment and family well-being. These surveys will also provide us with detailed information about the counterfactual condition, highlighting how young children in half-day programs spend the out-of-school portions of their day.

Finally, it is essential to consider the magnitude of the benefits observed from full-day pre-school in light of the program's costs, and compared to other, potentially less expensive approaches to supporting ECE programs. Moving from half- to full-day preschool is a relatively costly policy, because half as many students can be accommodated by the same number of classrooms and teachers (a full-day classroom accommodates 16 students each day, and a half-day classroom accommodates 32). An intensive cost-benefit analyses is beyond the scope of the current paper and planned for future work on this project. The various potential benefits, in particular, are difficult to monetize (Levin, McEwan, Belfield, Bowden, & Shand, 2017). However, to provide some sense of the cost side, WPS spends about $4,180 additional dollars per student to offer full-

day. In addition, districts in Colorado receive about $4,400 per child from the Colorado Department of Education, and if the district serves fewer students due to offering full-day classrooms, they receive less of this funding.

Policy Implications and Conclusions

While more research is certainly needed to examine exactly who benefits from more intensive ECE programs, on which outcomes, and through what mechanisms, the current study does provide the most compelling evidence available to date that a full-day, full-week preschool supports young children's development, at least among a sample of primarily low-income, Latinx children. Our findings, coupled with the very high demand for full-day slots in this district, suggest that policy initiatives that provide greater access to full-day programs may be beneficial.

In recent years, the rapid fade-out of ECE program effects documented in several rigorous studies has led policy-makers and researchers to ask how best to ensure that ECE programs yield meaningful and long-lasting effects. Many have suggested that focusing on children's subsequent experiences in early elementary school is a critical strategy to better sustain the gains (Philips et al., 2017). While the focus on sustaining environments is certainly worthy of further investigation, the current study also suggests the importance of also focusing on the preschool year itself, and strategies for making that experience as meaningful for young children as possible.

Through this deep-dive into the impacts of one particular feature of ECE—program intensity—as well as through similar undertakings about other potentially central ECE features such as curricula, professional development, etc., we will begin to provide policy-makers with the kind of evidence necessary to make smart decisions not about whether or not to offer ECE programs but about *how* to design policies that yield meaningful and sustained impacts.

Tables and Figures

Table 1. Pre-K Study Sample Descriptive Statistics

	Pre-K Study Sample			
	Mean	SD	N	% Missing
Baseline Demographics				
% White	0.12	--	226	0.0%
% Hispanic	0.74	--	226	0.0%
% Home Lang. Not English	0.49	--	214	5.3%
% Parent Ed > HS	0.37	--	226	0.0%
% Free-Lunch Elig	0.61	--	226	0.0%
% Unknown Lunch Status	0.22	--	226	0.0%
% Red-Lunch Elig	0.13	--	226	0.0%
% Male	0.49	--	226	0.0%
% with Fam History of Special Needs	0.17	--	226	0.0%
% with Low Language Development	0.23	--	226	0.0%
% with Low Social Development	0.37	--	226	0.0%
Child's Age (in yrs)	4.36	0.30	223	1.3%
Assessment Variables				
PPVT PK Fall Std. Score	85.7	30.5	215	4.9%
ESI-R PK Fall Total Score	19.0	6.7	215	4.9%
PPVT PK Spring Std. Score	96.2	19.1	200	11.5%
ESI-R PK Spring Total Score	20.6	5.1	212	6.2%
TS GOLD PK Fall Cognitive Score	568.6	55.9	179	20.8%
TS GOLD PK Fall Language Score	560.3	52.2	179	20.8%
TS GOLD PK Fall Literacy Score	563.5	44.2	179	20.8%
TS GOLD PK Fall Math Score	567.3	46.0	178	21.2%
TS GOLD PK Fall Phys. Dev. Score	564.8	44.6	179	20.8%
TS GOLD PK Fall Soc. Emot. Score	576.8	49.4	179	20.8%
TS GOLD PK Spring Overall Score	659.5	52.5	182	19.5%
TS GOLD PK Spring Cognitive Score	676.2	63.5	182	19.5%
TS GOLD PK Spring Language Score	656.5	63.1	182	19.5%
TS GOLD PK Spring Literacy Score	653.2	54.5	182	19.5%
TS GOLD PK Spring Math Score	656.4	54.7	182	19.5%
TS GOLD PK Spring Phys. Dev. Score	647.8	57.3	182	19.5%
TS GOLD PK Spring Soc. Emot. Score	666.8	59.7	182	19.5%
DIBELS K Fall Composite Score	22.3	20.4	139	38.5%

FN: Table 1 presents descriptive statistics for the study sample on variables collected by the WPS Early Childhood Center or the study team. 226 study children were randomized. Demographic and family history questions come from the general application for WPS preschool. To see exact wording of these questions, see the Footnote of Table 2. All test scores are presented in Table 1 in their original (raw) metric. In analyses, they are standardized (mean=0, SD=1). The TS GOLD Spring Overall score in Table 1 is the mean of the 6 subdomain unstandardized scores. In the main analyses, the TS GOLD Spring Overall score is the mean of the standardized *subdomain scores. To compare the study sample to WPS overall, we examine demographics from the Common Core of Data from NCES for 2015-16. The full district is 52% male, 72% FL-eligible, 11% RL-eligible, 77% Hispanic, and 1% Black (U.S. Department of Education, National Center for Education Statistics, Common Core of Data, Retrieved from http://nces.ed.gov/ccd/elsi/).*

Table 2. Teacher Reports of Typical Classroom Time Use on 9 Activity Types

Activity	Mean Hours/Wk		Mean Mins/Day		Percent of School Day	
	Full	**Half**	**Full**	**Half**	**Full**	**Half**
Reading & Language Arts	3.7	1.3	44.3	20.0	10.9	10.9
Mathematics	2.4	1.0	28.0	15.8	6.9	8.7
Social Studies & Science	1.4	0.9	16.8	14.2	4.3	7.9
Eating	4.3	1.1	51.7	16.8	13.0	9.2
Napping & Resting	5.8	0.0	69.2	0.0	17.3	0.0
Music, Art, Drama, & Dance	1.8	0.9	21.3	14.2	5.3	7.8
Unstructured Play	7.9	3.8	95.2	57.7	23.8	31.7
Structured Play	3.6	1.6	43.5	23.8	10.9	13.1
Transitions	2.5	1.3	30.0	19.2	7.6	10.6
Totals	33.5	12.1	400.0	181.7	100.0	100.0

Note . Values have been rounded, and column totals may contain rounding error.

Table 3. Baseline Covariate Balance, Expressed in Original Metrics and Standardized Cohen's D

Pre-Treatment Covariate	Treatment Mean	Control Mean	Raw Difference	Cohen's D	T-Statistic	P-Value	N
% White	11.4%	11.6%	-0.2%	-0.006	0.005	0.996	-- 226
% Hispanic	72.8%	75.0%	-2.2%	-0.050	0.070	0.944	-- 226
% Home Lang. Not English	48.2%	49.0%	-0.8%	-0.016	0.028	0.977	-- 214
% Parent Ed > HS	36.8%	37.5%	-0.7%	-0.014	0.023	0.981	-- 226
% Free-Lunch Elig	66.7%	54.5%	12.2%	0.244	0.431	0.667	-- 226
% Unknown Lunch Status	21.1%	22.3%	-1.3%	-0.030	0.037	0.970	-- 226
% Red-Lunch Elig	11.4%	14.3%	-2.9%	-0.082	0.070	0.944	-- 226
% Male	48.2%	49.1%	-0.9%	-0.017	0.032	0.975	-- 226
% with Fam History of Special Needs	17.5%	16.1%	1.5%	0.040	0.040	0.968	-- 226
% with Low Language Development	24.6%	21.4%	3.1%	0.076	0.097	0.923	-- 226
% with Low Social Development	37.7%	35.7%	2.0%	0.042	0.069	0.945	-- 226
Child's Age (in yrs)	4.34	4.38	-0.04	-0.124	0.923	0.357	-- 223
PPVT PK Fall Std. Score	0.027	-0.034	0.061	0.061	0.015	0.988	-- 215
ESI-R PK Fall Total Score	0.084	-0.093	0.178	0.178	0.199	0.843	-- 215

*FN: We use logistic regression for binary outcomes. Demographic and family history questions come from the general application for WPS preschool. Parents are asked to answer yes or no to the following questions: Q1: Has an immediate family member received Special Education services? Q2: Is your child in need of language development including, but not limited to, the ability to speak English? Q3: Does your child have problems with social situations? (+ for p<.10, * for p<.05, ** for p<.01, and *** for p<.001). We also conduct a multivariate regression model for all 14 dependent variables, predicted by treatment status: F(15, 204) = 1.28, Prob > F = 0.215)*

Table 4. Baseline Covariate Balance Comparison: Full Sample vs. Sample with End-of-Pre-Kindergarten Outcomes

	Treatment Group				Control Group				Cohen's D & Sig	
	Full Sample		Non-Attrited		Full Sample		Non-Attrited		Full	Non-Attrited
Pre-Treatment Covariate	Mean	(N)	Mean	(N)	Mean	(N)	Mean	(N)	**Sample**	**Sample**
% White	11.4%	(114)	12.1%	(107)	11.6%	(112)	9.4%	(96)	-0.006 --	0.095 --
% Hispanic	72.8%	(114)	72.9%	(107)	75.0%	(112)	77.1%	(96)	-0.050 --	-0.099 --
% Home Lang. Not English	48.2%	(112)	47.6%	(105)	49.0%	(102)	50.0%	(88)	-0.016 --	-0.047 --
% Parent Ed > HS	36.8%	(114)	35.5%	(107)	37.5%	(112)	42.7%	(96)	-0.014 --	-0.145 --
% Free-Lunch Elig	66.7%	(114)	65.4%	(107)	54.5%	(112)	57.3%	(96)	0.244 --	0.163 --
% Unknown Lunch Status	21.1%	(114)	21.5%	(107)	22.3%	(112)	17.7%	(96)	-0.030 --	0.099 --
% Red-Lunch Elig	11.4%	(114)	12.1%	(107)	14.3%	(112)	15.6%	(96)	-0.082 --	-0.095 --
% Male	48.2%	(114)	50.5%	(107)	49.1%	(112)	49.0%	(96)	-0.017 --	0.030 --
% with Fam History of Special Needs	17.5%	(114)	17.8%	(107)	16.1%	(112)	17.7%	(96)	0.040 --	0.001 --
% with Low Language Development	24.6%	(114)	24.3%	(107)	21.4%	(112)	22.9%	(96)	0.076 --	0.033 --
% with Low Social Development	37.7%	(114)	38.3%	(107)	35.7%	(112)	35.4%	(96)	0.042 --	0.060 --
Child's Age (in yrs)	4.34	(114)	4.35	(107)	4.38	(109)	4.36	(93)	-0.124 --	-0.031 --
PPVT PK Fall Std. Score	0.027	(111)	0.015	(107)	-0.034	(104)	-0.093	(92)	0.061 --	0.107 --
ESI-R PK Fall Total Score	0.084	(111)	0.103	(107)	-0.093	(104)	-0.050	(92)	0.178 --	0.153 --

*FN: We use logistic regression for binary outcomes. "Full Sample" refers all study participants originally assigned to each treatment group. The "Non-Attrited" sample refers to study participants who are observed with PPVT test scores in the spring of pre-k. (-- for not sig (p>0.10), + for p<.10, * for p<.05, ** for p<.01, and *** for p<.001). We also conduct a multivariate regression model for all 14 dependent variables, predicted by treatment status:*
For Full Study Sample: F(15, 204)= 1.28, Prob> F= 0.215.
Non-Attrition Sample: F(15, 190)= 1.25, Prob> F= 0.238..

Table 5. Primary Outcomes (End of Pre-K): Causal Effects of Full-Day Pre-kindergarten (ITT vs. CATE)

Intent-to-Treat Analysis

	PPVT End of Pre-K			***ESI-R End of Pre-K***		
	(M1)	***(M2)***	***(M3)***	***(M1)***	***(M2)***	***(M3)***
Assigned to Full	0.306 *	0.303 *	0.275 **	0.140	0.142	0.101
	(0.136)	(0.127)	(0.087)	(0.143)	(0.146)	(0.133)
Constant	0.004	0.079	0.087	0.013	0.069	0.056
	(0.099)	(0.093)	(0.064)	(0.103)	(0.107)	(0.097)
R^2	0.055	0.300	0.680	0.026	0.136	0.310
Adj. R^2	0.020	0.230	0.642	-0.009	0.051	0.230
N	200	200	200	202	202	202

Complier Average Treatment Effects (Two-Stage Least Squares)

	PPVT End of Pre-K			***ESI-R End of Pre-K***		
	(M1)	***(M2)***	***(M3)***	***(M1)***	***(M2)***	***(M3)***
Attended Full-Day	0.399 *	0.397 *	0.363 **	0.174	0.185	0.132
	(0.172)	(0.166)	(0.116)	(0.178)	(0.189)	(0.172)
Constant	-0.044	0.025	0.040	-0.005	0.044	0.039
	(0.113)	(0.111)	(0.077)	(0.117)	(0.127)	(0.115)
N	200	200	200	202	202	202

*FN: M1 includes first-choice school (i.e., block) fixed effects only. In M2, we add student-level demographic control variables (the variables in Table 4, except baseline PPVT & ESI-R scores). In M3, we also add baseline PPVT & ESI-R scores). We include missingness dummies in cases where respondents have missing pre-treatment covariates. For the two-stage least squares analysis (lower panel), the F-statistic for the first stage equation from (M1) is 311.7 (+ for p<.10, * for p<.05, ** for p<.01, and *** for p<.001)*

Table 6. Exploratory Outcomes: Causal Effects of Full-Day Pre-kindergarten (ITT vs. CATE), Full Model (3) Only

	End of Pre-K TS GOLD							Fall Kindergarten DIBELS		
	Overall	*Cognition*	*Language*	*Literacy*	*Math*	*Physical*	*Socio-Emotional*	*Overall Composite*	*1st Sound Fluency*	*Ltr Naming Fluency*
Assigned to Full	0.253 ***	0.258 **	0.114	0.393 ***	0.230 **	0.237 **	0.153 *	0.344 *	0.266	0.354 *
	(0.073)	(0.087)	(0.084)	(0.084)	(0.081)	(0.081)	(0.074)	(0.167)	(0.193)	(0.172)
Constant	-0.008	0.013	0.022	0.028	0.010	-0.032	-0.078	0.143	0.118	0.141
	(0.053)	(0.064)	(0.062)	(0.062)	(0.059)	(0.060)	(0.054)	(0.128)	(0.147)	(0.131)
R^2	0.812	0.722	0.735	0.795	0.789	0.753	0.790	0.566	0.475	0.515
Adj. R^2	0.770	0.660	0.676	0.749	0.742	0.698	0.743	0.435	0.317	0.369
N	182	182	182	182	182	182	182	139	139	139

Complier Average Treatment Effects (Two-Stage Least Squares)

	End of Pre-K TS GOLD							Fall Kindergarten DIBELS		
	Overall	*Cognition*	*Language*	*Literacy*	*Math*	*Physical*	*Socio-Emotional*	*Overall Composite*	*1st Sound Fluency*	*Ltr Naming Fluency*
Attended Full-Day	0.314 ***	0.320 **	0.141	0.487 ***	0.285 **	0.294 **	0.190 *	0.392 *	0.303	0.404 *
	(0.088)	(0.106)	(0.104)	(0.101)	(0.098)	(0.102)	(0.091)	(0.191)	(0.219)	(0.196)
Constant	-0.055	-0.035	0.001	-0.045	-0.032	-0.076	-0.106	0.112	0.094	0.109
	(0.063)	(0.076)	(0.074)	(0.073)	(0.070)	(0.073)	(0.065)	(0.140)	(0.161)	(0.144)
N	182	182	182	182	182	182	182	139	139	139

*FN: Results are reported for M3 only, which includes first-choice school (i.e., block) fixed effects, student-level demographic control variables, and baseline PPVT, ESI-R, and TS GOLD scores. Following the practice of Russo et al. (2019), we produce an End of Pre-K TS GOLD Overall score by taking the mean of the 6 standardized subdomain scores. We include missingness dummies in cases where respondents have missing pre-treatment covariates. For the two-stage least squares analysis (lower panel), the F-statistic for the first stage equation from (M1) is 311.7 (+ for p<.10, * for p<.05, ** for p<.01, and *** for p<.001)*

Table 7. Robustness Check: Causal Effects of Full-Day Pre-K on All Outcomes, when Missing Outcomes Imputed (ITT vs. CATE).

Intent-to-Treat Analysis

	Primary Outcomes		End of Pre-K TS GOLD							Fall Kindergarten DIBELS		
	PPVT	*ESI-R*	*Overall*	*Cognition*	*Language*	*Literacy*	*Math*	*Physical*	*Socio-Emotional*	*Overall Composite*	*1st Sound Fluency*	*Ltr Naming Fluency*
Assigned to Full	0.220 *	0.036	0.201 **	0.217 **	0.083	0.267 **	0.176 *	0.191 **	0.153 *	0.093	0.066	0.109
	(0.085)	(0.124)	(0.068)	(0.080)	(0.076)	(0.080)	(0.075)	(0.072)	(0.066)	(0.130)	(0.143)	(0.131)
Constant	0.036	0.058	-0.034	-0.011	-0.003	0.002	-0.041	-0.052	-0.072	0.125	0.099	0.129
	(0.061)	(0.090)	(0.052)	(0.061)	(0.058)	(0.061)	(0.057)	(0.055)	(0.051)	(0.099)	(0.109)	(0.100)
R^2	0.612	0.230	0.762	0.662	0.690	0.718	0.726	0.709	0.759	0.344	0.281	0.307
Adj. R^2	0.594	0.194	0.742	0.632	0.663	0.693	0.703	0.684	0.738	0.287	0.219	0.247
N	226	226	226	226	226	226	226	226	226	226	226	226

Complier Average Treatment Effects (Two-Stage Least Squares)

	Primary Outcomes		End of Pre-K TS GOLD							Fall Kindergarten DIBELS		
	PPVT	*ESI-R*	*Overall*	*Cognition*	*Language*	*Literacy*	*Math*	*Physical*	*Socio-Emotional*	*Overall Composite*	*1st Sound Fluency*	*Ltr Naming Fluency*
Attended Full-Day	0.297 **	0.048	0.309 **	0.333 **	0.127	0.411 ***	0.270 *	0.294 **	0.235 *	0.117	0.075	0.141
	(0.114)	(0.168)	(0.100)	(0.118)	(0.116)	(0.116)	(0.112)	(0.110)	(0.100)	(0.199)	(0.219)	(0.201)
Constant	0.001	0.052	-0.107	-0.090	-0.033	-0.095	-0.104	-0.121	-0.127 +	0.109	0.093	0.107
	(0.071)	(0.105)	(0.069)	(0.081)	(0.080)	(0.080)	(0.077)	(0.076)	(0.069)	(0.137)	(0.151)	(0.139)
N	226	226	226	226	226	226	226	226	226	226	226	226

*FN: Results are reported for M3 only, which includes first-choice school (i.e., block) fixed effects, student-level demographic control variables, and baseline PPVT, ESI-R, and TS GOLD scores. Following the practice of Russo et al. (2019), we produce an End of Pre-K TS GOLD Overall score by taking the mean of the 6 standardized subdomain scores. We include missingness dummies in cases where respondents have missing pre-treatment covariates. (+ for p<.10, * for p<.05, ** for p<.01, and *** for p<.001)*

Appendix A: Baseline Covariate Balance and Attrition

Table A1. Baseline Covariate Balance Comparison: Full Sample vs. Sample with End-of-Preschool TS-GOLD Exploratory Outcomes (Administered by District)

	Treatment Group				Control Group				Cohen's D & Sig Stars	
	Full Sample		Non-Attrited		Full Sample		Non-Attrited		Full Sample	Non-Attrited
Pre-Treatment Covariate	Mean	(N)	Mean	(N)	Mean	(N)	Mean	(N)	D SIG	D SIG
% White	11.4%	(114)	11.9%	(101)	11.6%	(112)	12.3%	(81)	-0.006 --	-0.014 --
% Hispanic	72.8%	(114)	73.3%	(101)	75.0%	(112)	79.0%	(81)	-0.050 --	-0.140 --
% Home Lang. Not English	48.2%	(112)	48.5%	(101)	49.0%	(102)	48.1%	(79)	-0.016 --	0.008 --
% Parent Ed > HS	36.8%	(114)	37.6%	(101)	37.5%	(112)	46.9%	(81)	-0.014 --	-0.185 --
% Free-Lunch Elig	66.7%	(114)	63.4%	(101)	54.5%	(112)	61.7%	(81)	0.244 --	0.033 --
% Unknown Lunch Status	21.1%	(114)	22.8%	(101)	22.3%	(112)	11.1%	(81)	-0.030 --	0.369 --
% Red-Lunch Elig	11.4%	(114)	12.9%	(101)	14.3%	(112)	18.5%	(81)	-0.082 --	-0.144 --
% Male	48.2%	(114)	49.5%	(101)	49.1%	(112)	51.9%	(81)	-0.017 --	-0.047 --
% with Fam History of Special Needs	17.5%	(114)	18.8%	(101)	16.1%	(112)	19.8%	(81)	0.040 --	-0.023 --
% with Low Language Development	24.6%	(114)	25.7%	(101)	21.4%	(112)	24.7%	(81)	0.076 --	0.024 --
% with Low Social Development	37.7%	(114)	37.6%	(101)	35.7%	(112)	39.5%	(81)	0.042 --	-0.038 --
Child's Age (in yrs)	4.34	(114)	4.36	(101)	4.38	(109)	4.38	(80)	-0.124 --	-0.066 --
PPVT PK Fall Std. Score	0.027	(111)	0.032	(101)	-0.034	(104)	-0.029	(79)	0.061 --	0.060 --
ESI-R PK Fall Total Score	0.084	(111)	0.140	(101)	-0.093	(104)	-0.111	(79)	0.178 --	0.251 --

*FN: We use logistic regression for binary outcomes. "Full Sample" refers all study participants originally assigned to each treatment group. The "Non-Attrited" sample refers to study participants who are observed with TS GOLD test scores in the spring of pre-k. (-- for not sig (p>0.10), + for p<.10, * for p<.05, ** for p<.01, and *** for p<.001). We also conduct a multivariate regression model for all 14 dependent variables, predicted by treatment status. For Full Study Sample: F(15, 204)= 1.28, Prob> F= 0.215. For Non-Attrition Study Sample: F(15, 177) = 1.37, Prob > F = 0.169)*

Table A2. Baseline Covariate Balance Comparison: Full Sample vs. Sample with Fall of Kindergarten DIBELS Exploratory Outcomes (Administered by District)

	Treatment Group				Control Group				Cohen's D & Sig Stars	
	Full Sample		Non-Attrited		Full Sample		Non-Attrited		Full Sample	Non-Attrited
Pre-Treatment Covariate	Mean	(N)	Mean	(N)	Mean	(N)	Mean	(N)	D SIG	D SIG
% White	11.4%	(114)	10.3%	(78)	11.6%	(112)	11.5%	(61)	-0.006 --	-0.038 --
% Hispanic	72.8%	(114)	75.6%	(78)	75.0%	(112)	78.7%	(61)	-0.050 --	-0.074 --
% Home Lang. Not English	48.2%	(112)	47.4%	(78)	49.0%	(102)	51.7%	(58)	-0.016 --	-0.085 --
% Parent Ed > HS	36.8%	(114)	35.9%	(78)	37.5%	(112)	39.3%	(61)	-0.014 --	-0.070 --
% Free-Lunch Elig	66.7%	(114)	62.8%	(78)	54.5%	(112)	55.7%	(61)	0.244 --	0.141 --
% Unknown Lunch Status	21.1%	(114)	23.1%	(78)	22.3%	(112)	18.0%	(61)	-0.030 --	0.130 --
% Red-Lunch Elig	11.4%	(114)	12.8%	(78)	14.3%	(112)	19.7%	(61)	-0.082 --	-0.171 --
% Male	48.2%	(114)	48.7%	(78)	49.1%	(112)	54.1%	(61)	-0.017 --	-0.107 --
% with Fam History of Special Needs	17.5%	(114)	21.8%	(78)	16.1%	(112)	13.1%	(61)	0.040 --	0.245 --
% with Low Language Development	24.6%	(114)	21.8%	(78)	21.4%	(112)	24.6%	(61)	0.076 --	-0.064 --
% with Low Social Development	37.7%	(114)	34.6%	(78)	35.7%	(112)	39.3%	(61)	0.042 --	-0.096 --
Child's Age (in yrs)	4.34	(114)	4.34	(78)	4.38	(109)	4.33	(59)	-0.124 --	0.019 --
PPVT PK Fall Std. Score	0.027	(111)	0.072	(78)	-0.034	(104)	-0.193	(59)	0.061 --	0.215 --
ESI-R PK Fall Total Score	0.084	(111)	0.211	(78)	-0.093	(104)	-0.002	(59)	0.178 --	0.213 --

*FN: We use logistic regression for binary outcomes. "Full Sample" refers all study participants originally assigned to each treatment group. The "Non-Attrited" sample refers to study participants who are observed with DIBELS test scores in the fall of k. (-- for not sig (p>0.10), + for p<.10, * for p<.05, ** for p<.01, and *** for p<.001). We conduct a multivariate regression model for all 14 dependent variables, predicted by treatment status. For Full Study Sample: F(15, 204) = 1.28, Prob > F = 0.215), For Non-Attrition Study Sample: F(15, 134) = 1.03, Prob > F = 0.430).*

References

Abry, T., Latham, S., Bassok, D., & LoCasale-Crouch, J. (2015). Preschool and kindergarten teachers' beliefs about early school competencies: Misalignment matters for kindergarten adjustment. *Early Childhood Research Quarterly, 31*, 78-88.

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., & Schady, N. (2016). Teacher quality and learning outcomes in kindergarten. *The Quarterly Journal of Economics, 131*(3), 1415-1453.

Barnett, W. S., Friedman-Krauss, A., Weisenfeld, G., Horowitz, M., Kasmin, R., & Squires, J. (2017). The state of preschool 2016: State preschool yearbook. New Brunswick, NJ: National Institute for Early Education Research: National Institute for Early Education Research New Brunswick, NJ.

Barnett, W. S., Jung, K., Friedman-Krauss, A., Frede, E. C., Nores, M., Hustedt, J. T., . . . Daniel-Echols, M. (2018). State prekindergarten effects on early learning at kindergarten entry: An analysis of eight state programs. *AERA Open, 4*(2), 2332858418766291.

Bassok, D. (2010). Do Black and Hispanic children benefit more from preschool? Understanding differences in preschool effects across racial groups. *Child Development, 81*(6), 1828-1845.

Bassok, D., Fitzpatrick, M., Greenberg, E., & Loeb, S. (2016). Within-and between-sector quality differences in early childhood education and care. *Child Development, 87*(5), 1627-1645.

Bassok, D., Gibbs, C. R., & Latham, S. (2018). Preschool and Children's Outcomes in Elementary School: Have Patterns Changed Nationwide Between 1998 and 2010? *Child Development*.

Battistin, E., & Meroni, E. C. (2016). Should we increase instruction time in low achieving schools? Evidence from Southern Italy. *Economics of Education Review, 55*, 39-56.

Bauer, L., & Schanzenbach, D. W. (2016). The long-term impact of the Head Start program. *The Hamilton Project*.

Bauernschuster, S., & Schlotter, M. (2015). Public child care and mothers' labor supply—Evidence from two quasi-experiments. *Journal of Public Economics, 123*, 1-16.

Bellei, C. (2009). Does lengthening the school day increase students' academic achievement? Results from a natural experiment in Chile. *Economics of Education Review, 28*(5), 629-640.

Belsky, J. (2002). Quantity counts: Amount of child care and children's socioemotional development. *Journal of Developmental & Behavioral Pediatrics, 23*(3), 167-170.

Blau, D., & Currie, J. (2006). Pre-school, day care, and after-school care: who's minding the kids? *Handbook of the Economics of Education, 2*, 1163-1278.

Brownell, M. D., Nickel, N. C., Chateau, D., Martens, P. J., Taylor, C., Crockett, L., . . . Goh, C. Y. (2015). Long-term benefits of full-day kindergarten: a longitudinal population-based study. *Early Child Development and Care, 185*(2), 291-316.

Burke, M. D., Hagan-Burke, S., Kwok, O., & Parker, R. (2009). Predictive validity of early literacy indicators from the middle of kindergarten to second grade. *The Journal of Special Education, 42*(4), 209-226.

Campbell, F. A., Pungello, E. P., Burchinal, M., Kainz, K., Pan, Y., Wasik, B. H., . . . Ramey, C. T. (2012). Adult outcomes as a function of an early childhood educational program: an Abecedarian Project follow-up. *Developmental psychology, 48*(4), 1033.

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

Cascio, E. U. (2009). Maternal labor supply and the introduction of kindergartens into American public schools. *Journal of Human Resources, 44*(1), 140-170.

Clements, D. H., & Sarama, J. (2008). Experimental evaluation of the effects of a research-based preschool mathematics curriculum. *American Educational Research Journal, 45*(2), 443-494.

Cooper, H., Allen, A. B., Patall, E. A., & Dent, A. L. (2010). Effects of full-day kindergarten on academic achievement and social development. *Review of Educational Research, 80*(1), 34-70.

Corcoran, R. P., Eisinger, J., Kim, E., & , & Ross, S. M. (2016). *An evaluation of the McGraw-Hill education reading wonders program*. Retrieved from Center for Research & Reform in Education; Corcoran Lab.: https://s3.amazonaws.com/ecommerce-prod.mheducation.com/unitas/school/explore/sites/reading-wonders/wonders-research-evidence-compendium.pdf

Curenton, S. M. (2011). Understanding the landscapes of stories: The association between preschoolers' narrative comprehension and production skills and cognitive abilities. *Early Child Development and Care, 181*(6), 791-808.

Duncan, G. J., Engel, M., Claessens, A., & Dowsett, C. J. (2014). Replication and robustness in developmental research. *Developmental psychology, 50*(11), 2417.

Dunn, L., & Dunn, D. (2013). PPVT-4 Technical Report: Minneapolis, MN: Pearson Assessments.

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

Early, D. M., Maxwell, K. L., Ponder, B. D., & Pan, Y. (2017). Improving teacher-child interactions: A randomized controlled trial of Making the Most of Classroom Interactions and My Teaching Partner professional development models. *Early Childhood Research Quarterly, 38*, 57-70.

Fantuzzo, J., Perry, M. A., & McDermott, P. (2004). Preschool approaches to learning and their relationship to other relevant classroom competencies for low-income children. *School Psychology Quarterly, 19*(3), 212.

Figlio, D., Holden, K. L., & Ozek, U. (2018). Do students benefit from longer school days? Regression discontinuity evidence from Florida's additional hour of literacy instruction. *Economics of Education Review, 67*, 171-183.

Gelber, A., & Isen, A. (2013). Children's schooling and parents' behavior: Evidence from the Head Start Impact Study. *Journal of Public Economics, 101*, 25-38.

Gibbs, C. (2014a). Experimental evidence on early intervention: The impact of full-day kindergarten. *Frank Batten School of Leadership and Public Policy Working Paper, 4*.

Gibbs, C. (2014b). Experimental evidence on early intervention: The impact of full-day kindergarten. *Batten School of Leadership and Public Policy, University of Virginia http://www. batten. virginia. edu/sites/default/files/fwpapers/Gibbs_full-day% 20K% 20experiment. pdf*.

Good, R., & Kaminski, R. (2002). *Dynamic Indicators of Basic Early Literacy Skills (DIBELS)* (Vol. 6). Eugene, OR: Institute for the Development of Education Achievement. Available: http://dibels.uoregon.edu.

Good, R., Kaminski, R., Shinn, M., Bratten, J., Shinn, M., Laimon, D., & Flindt, N. (2004). Technical adequacy of DIBELS: Results of the Early Childhood Research Institute on measuring growth and development. *Eugene, OR: University of Oregon*.

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.

Gormley, W. T. (2008). The effects of Oklahoma's pre-k program on Hispanic children. *Social Science Quarterly, 89*(4), 916-936.

Gullo, D. F. (2000). The long term educational effects of half-day vs full-day kindergarten. *Early Child Development and Care, 160*(1), 17-24.

Head Start Performance Standards. (2016). Washington, DC: Federal Register Vol 81, Issue 172, Sept 6, 2016 Retrieved from https://www.gpo.gov/fdsys/pkg/FR-2016-09-06/pdf/2016-19748.pdf.

Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science, 312*(5782), 1900-1902.

Herbst, C. M. (2017). Universal child care, maternal employment, and children's long-run outcomes: Evidence from the US Lanham Act of 1940. *Journal of Labor Economics, 35*(2), 519-564.

Heroman, C., Burts, D. C., Berke, K.-l., Bickart, T. S., Nelson, H. P., Taub, L., & Boyle, K. (2010). Objectives for Development & Learning: Birth Through Kindergarten.

Herry, Y., Maltais, C., & Thompson, K. (2007). Effects of a full-day preschool program on 4-year-old children. *Early Childhood Research & Practice, 9*(2), n2.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis, 15*(3), 199-236.

Interview With Early Childhood Department Leadership (2016). [Prediscussion of RCT study Design with WPS ECC Leadership].

Kena, G., Hussar, W., McFarland, J., de Brey, C., Musu-Gillette, L., Wang, X., . . . Diliberti, M. (2016). The Condition of Education 2016. NCES 2016-144. *National Center for Education Statistics*.

Kim, D.-H., Lambert, R. G., & Burts, D. C. (2013). Evidence of the validity of Teaching Strategies GOLD® assessment tool for English language learners and children with disabilities. *Early Education & Development, 24*(4), 574-595.

Leow, C., & Wen, X. (2017). Is full day better than half day? A propensity score analysis of the association between Head Start Program intensity and children's school performance in kindergarten. *Early Education and Development, 28*(2), 224-239.

Levin, H. M., McEwan, P. J., Belfield, C., Bowden, A. B., & Shand, R. (2017). *Economic evaluation in education: Cost-effectiveness and benefit-cost analysis*: SAGE publications.

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.

Loeb, S., Bridges, M., Bassok, D., Fuller, B., & Rumberger, R. W. (2007). How much is too much? The influence of preschool centers on children's social and cognitive development. *Economics of Education Review, 26*(1), 52-66.

Luo, Z., Jose, P. E., Huntsinger, C. S., & Pigott, T. D. (2007). Fine motor skills and mathematics achievement in East Asian American and European American kindergartners and first graders. *British Journal of Developmental Psychology, 25*(4), 595-614.

Macmillan/McGraw-Hill. (2019). Welcome to MacMillan McGraw-Hill's Little Treasures. Retrieved from http://reading.macmillanmh.com/ltreasures/

Malik, R. (2018). The Effects of Universal Preschool in Washington, DC. *Washington: Center for American Progress*.

Meisels, S. J., Henderson, L. W., Liaw, F.-r., Browning, K., & Ten Have, T. (1993). New evidence for the effectiveness of the Early Screening Inventory. *Early Childhood Research Quarterly, 8*(3), 327-346.

Meisels, S. J., Marsden, D. B., Wiske, M. S., & Henderson, L. W. (1997). *ESI-R: Early Screening Inventory-Revised. Examiner's Manual*: ERIC.

Miller-Bains, K. L., Russo, J. M., Williford, A. P., DeCoster, J., & Cottone, E. A. (2017). Examining the validity of a multidimensional performance-based assessment at kindergarten entry. *AERA Open, 3*(2), 2332858417706969.

Moodie, S., Daneri, P., Goldhagen, S., Halle, T., Green, K., & LaMonte, L. (2014). Early Childhood Developmental Screening: A Compendium of Measures for Children Ages Birth to Five. OPRE Report 2014-11. *US Department of Health and Human Services*.

Morris, P. A., Connors, M., Friedman-Krauss, A., McCoy, D. C., Weiland, C., Feller, A., . . . Yoshikawa, H. (2018). New findings on impact variation from the Head Start Impact Study: Informing the scale-up of early childhood programs. *AERA Open, 4*(2), 2332858418769287.

Morrissey, T. W. (2009). Multiple child-care arrangements and young children's behavioral outcomes. *Child Development, 80*(1), 59-76.

Morrissey, T. W. (2013). Multiple child care arrangements and common communicable illnesses in children aged 3 to 54 months. *Maternal and child health journal, 17*(7), 1175-1184.

Philips, D. A., Lipsey, M. W., Dodge, K. A., Haskins, R., Bassok, D., Burchinal, M., . . . Weiland, C. (2017). *Puzzling It Out: The Current State of Scientific Knowledge on Pre-Kindergarten Effects: A Consensus Statement*. Retrieved from https://www.brookings.edu/wp-content/uploads/2017/04/consensus-statement_final.pdf

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

Piasta, S. B., Justice, L. M., O'Connell, A. A., Mauck, S. A., Weber-Mayrer, M., Schachter, R. E., . . . Spear, C. F. (2017). Effectiveness of large-scale, state-sponsored language and literacy professional development on early childhood educator outcomes. *Journal of Research on Educational Effectiveness, 10*(2), 354-378.

Pilarz, A. R., & Hill, H. D. (2014). Unstable and multiple child care arrangements and young children's behavior. *Early Childhood Research Quarterly, 29*(4), 471-483.

Puma, M., Bell, S., Cook, R., Heid, C., Shapiro, G., Broene, P., . . . Friedman, J. (2010). Head Start Impact Study. Final Report. *Administration for Children & Families*.

Reynolds, A. J., Richardson, B. A., Hayakawa, M., Lease, E. M., Warner-Richter, M., Englund, M. M., . . . Sullivan, M. (2014). Association of a full-day vs part-day preschool intervention with school readiness, attendance, and parent involvement. *Jama, 312*(20), 2126-2134.

Robin, K. B., Frede, E. C., & Barnett, W. S. (2006). Is more better? The effects of full-day vs half-day preschool on early school achievement.

Rock, D. A., & Pollack, J. M. (2002). Early Childhood Longitudinal Study-Kindergarten Class of 1998-99 (ECLS-K): Psychometric Report for Kindergarten through First Grade. Working Paper Series.

Rouse, H. L., & Fantuzzo, J. (2006). Validity of the Dynamic Indicators for Basic Early Literacy Skills as an indicator of early literacy for urban kindergarten children. *School Psychology Review, 35*(3), 341.

Russo, J. M., Williford, A. P., Markowitz, A. J., Vitiello, V. E., & Bassok, D. (2019). Examining the validity of a widely-used school readiness assessment: Implications for teachers and early childhood programs. *Early Childhood Research Quarterly, 48*, 14-25.

Sabol, T. J., & Chase-Lansdale, P. L. (2015). The influence of low-income children's participation in Head Start on their parents' education and employment. *Journal of Policy Analysis and Management, 34*(1), 136-161.

Shah, H. K., Domitrovich, C. E., Morgan, N. R., Moore, J. E., Cooper, B. R., Jacobson, L., & Greenberg, M. T. (2017). One or two years of participation: Is dosage of an enhanced publicly funded preschool program associated with the academic and executive function skills of low-income children in early elementary school? *Early Childhood Research Quarterly, 40*, 123-137.

Shonkoff, J. P., & Phillips, D. A. (2000). *From neurons to neighborhoods: The science of early childhood development*: National Academies Press.

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

Teaching Strategies. (2011). *Teaching Strategies GOLD Assessment System Technical Summary*. Retrieved from

Teaching Strategies. (2013). *Teaching strategies GOLD assessment system: Concurrent validity*. Retrieved from

Vandell, D. L., Belsky, J., Burchinal, M., Steinberg, L., & Vandergrift, N. (2010). Do effects of early child care extend to age 15 years? Results from the NICHD study of early child care and youth development. *Child Development, 81*(3), 737-756.

Walters, C. R. (2015). Inputs in the production of early childhood human capital: Evidence from Head Start. *American Economic Journal: Applied Economics, 7*(4), 76-102.

Weiland, C. (2018). Pivoting to the "how": Moving preschool policy, practice, and research forward. *Early Childhood Research Quarterly*.

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.

Wen, X., Leow, C., Hahs-Vaughn, D. L., Korfmacher, J., & Marcus, S. M. (2012). Are two years better than one year? A propensity score analysis of the impact of Head Start program duration on children's school performance in kindergarten. *Early Childhood Research Quarterly, 27*(4), 684-694.

What Works Clearinghouse. (2014). Procedures and standards handbook (Version 3.0). *Washington, DC: US Department of Education*.

Whitebook, M., Phillips, D., & Howes, C. (2014). Worthy work, STILL unlivable wages: The early childhood workforce 25 years after the National Child Care Staffing Study. *Center for the Study of Child Care Employment, University of California, Berkeley*.

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

WPS TCAP Results. (2014). *Comparison of Third Grade TCAP Math Results for 2005-2014: Adams County School District 50*. Retrieved from Westminster Public School District: https://www.westminsterpublicschools.org/Page/15

Zvoch, K., Reynolds, R. E., & Parker, R. P. (2008). Full-day kindergarten and student literacy growth: Does a lengthened school day make a difference? *Early Childhood Research Quarterly*, *23*(1), 94-107.