Early Childhood Research Quarterly 27 (2012) 684–694

Contents lists available at ScienceDirect

Early Childhood Research Quarterly

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[☆]

Xiaoli Wen[a,*], Christine Leow[b], Debbie L. Hahs-Vaughn[c], Jon Korfmacher[b], Sue M. Marcus[d]

[a] *National-Louis University, Early Childhood Education Department, 122 South Michigan Avenue, Chicago, IL 60603, United States*
[b] *Erikson Institute, United States*
[c] *University of Central Florida, United States*
[d] *Columbia University, United States*

ARTICLE INFO

Keywords:
Head Start
Program duration
School performance
Propensity scores

ABSTRACT

Using data from a nationally representative sample, this study examined Head Start children's school outcome differences by the end of Kindergarten between children who attended Head Start program for two years and the ones who attended for one year. Propensity scores were used to match children who experienced different durations of the program on a series of demographic characteristics in order to achieve a precise estimation of the effects of program duration. The results showed that in comparison to a demographically comparable group of children who attended the Head Start program for one year, the children who experienced two years of intervention services had statistically significantly higher performance on all six academic and social outcome measures by the end of Kindergarten, which included PPVT, Woodcock–Johnson Reading Skills, Woodcock–Johnson Math Reasoning Skills, teacher-reported composite academic skills, preschool learning behaviors, and social skills. Policy and practice implications are discussed.

Early educational interventions, such as Head Start, have been widely recognized as an effective way to mitigate the negative effects of poverty on early learning and development (Burger, 2010; Camilli, Vargas, Ryan, & Barnett, 2010). Participation in early education interventions has shown short- and long-term positive effects on low-income children's academic skills, language development, social competence, emotional adjustment, reduced grade retention, and reduced need for special education (e.g., Belfield, Nores, Barnett, & Schweinhart, 2006; Gory, 2001; Ludwig & Phillips, 2007; Magnuson, Ruhm, & Waldfogel, 2007; Ramey et al., 2000; Temple & Reynolds, 2007). But what we know about early education interventions has primarily focused on the overall effectiveness, and we know much less about the specific program and participant factors and mechanisms that lead to favorable program outcomes (Berlin, O'Neal, & Brooks-Gunn, 1998; Guralnick, 1997; Reynolds, 2004). In this study, we looked at Head Start programs, the nation's largest early educational intervention, and examined the impact of one program design factor, program duration (defined as the length of program enrollment), on child outcomes.

Program duration is one way of measuring dosage, or the amount of intervention services families received. Intervention dosage is a multi-dimensional construct, and has been measured in several different forms beyond duration, including the amount of program contact (e.g., number of activities attended), intensity of intervention (e.g., full-day vs. half-day), and percentage or ratio of completed to expected amount of program contact, as defined by program protocol (Korfmacher et al., 2008; Littell, Alexander, & Reynolds, 2001). This study focuses on program duration in center-based educational intervention programs for low-income children because of its significant policy implications for the field of early education and intervention. In the past decade, there has been a strong expansion of early childhood programming (including Head Start and state-funded prekindergarten programs), but recent economic uncertainty calls into question the extent to which this expansion can be maintained. A tension exists between serving as many children as possible and providing the most impact with limited economic resources (e.g., Barnett & Hustedt, 2011; Steuerle, Reynolds, & Carasso, 2007), making the study of program design factors, such as length of programming, critical to efforts to serve low-income or at-risk children in the most efficient fashion.

Does the amount of intervention services children receive have causal impact on the amount of gains they accrue from the program? Theoretically and hypothetically, the dosage of intervention has been highlighted as a key variable in predicting program effectiveness (Shonkoff & Phillips, 2000). Available research evidence

[☆] The authors thank the Inter-University Consortium for Political and Social Research (ICPSR) for providing access to the FACES 2003 database. We are especially grateful to the children, parents, and teachers who participated in this study.

* Corresponding author. Tel.: +1 312 261 3160; fax: +1 312 261 3160.
E-mail address: Xiaoli.Wen@nl.edu (X. Wen).

0885-2006/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ecresq.2011.07.006

suggests that the most effective early intervention programs that were able to maintain long-term impacts were those that begin during children's early years of life, continue for multiple years, and provide support to families (e.g., Bogard & Takanishi, 2005; Reynolds, Ou, & Topitzes, 2004; Zigler & Styfco, 1994). Therefore, the length of intervention might be at least one of the factors that could determine the magnitude of program impact. However, it is challenging to make a causal conclusion regarding whether children and families who experience a longer duration of intervention would perform better on measured program outcomes than those who are enrolled for a relatively shorter time, because participants who experienced different amounts of intervention may differ in other ways as well, including their demographic characteristics (Hill, Brooks-Gunn, & Waldfogel, 2003; Powell, 2005). Simply stratifying participants by intervention duration or estimating the impact of duration in a standard regression model will not typically yield unbiased estimates because selection bias might be operating.

In the current study, we used a rigorous statistical methodology to make a less biased estimation of the impact of Head Start participation duration on Head Start children's school performance by the end of kindergarten. This study utilized secondary data from the national evaluation of Head Start, known as the Family and Child Experiences Survey (FACES, 2003 cohort; U.S. DHHS, 2008), to address whether Head Start children who entered the program at three years of age and were eligible to receive two years of program services performed academically and socially better in kindergarten than those who entered the program at four years of age and were eligible to receive only one year of program services.

1. Program duration impact

Intervention duration, as a critical program design component, has not received much research attention (Reynolds, 2004). A meta-analysis of 123 comparative studies of preschool intervention programs concluded that most studies did not collect detailed information on program duration, and for those studies that did examine duration, no significant impact on child cognitive and social outcomes was found (Camilli et al., 2010). Most empirical evidence regarding the impact of program duration is from examinations of small-scale model programs that usually are initiated by researchers, operate on a single site, have relatively narrow program foci, and are subject to close quality monitoring.

As one example, studies of the Carolina Abecedarian Project provide strong evidence regarding the effects of intervention duration (Campbell & Ramey, 1994). The Carolina Abecedarian Project is an early educational intervention for impoverished children, designed to prevent mild retardation and school failure through the provision of a supportive learning environment, beginning in infancy. In an experimental study, children were randomly assigned to one of the four conditions: educational treatment from infancy through third grade in public school (i.e., up to age eight); infancy through preschool treatment only (i.e., infancy to age five); primary school treatment only (ages five to eight); or an untreated control group. The design permitted the investigators to estimate the relative efficacy of interventions with different durations and different entry ages. The study also allows the estimation of the importance of reinforcing children's gains from early childhood treatment during the transition into early elementary school. A follow-up study of the children of the Carolina Abecedarian Project four to seven years after the intervention showed that the length of duration predicted Verbal IQ, Reading, and achievement scores. Findings generally supported the hypothesis that child academic performance increased as duration of treatment increased (eight years > five years > three years > none) (Campbell & Ramey, 1994). A more extended follow-up study of the same group of children when they reached age 15 demonstrated a similar conclusion: the largest program effects accrued to children whose program participation continued the longest (Campbell & Ramey, 1995).

Demonstration programs such as Abecedarian are different from large public programs sponsored by federal or state governments, and one must be cautious in generalizing findings that emerge from these model programs to large-scale public programs such as Head Start. The link between program duration and outcomes is not well established in Head Start programs. One study with a small sample of Head Start children explicitly compared child and family outcomes between children who attended the program for one year and those who attended for two years, and found that program duration was positively associated with home environment and parents' frequency of reading to children, but not child outcomes (Ritblatt, Brassert, Johnson, & Gomez, 2001). This study of Head Start population is correlational. Although attempts were made to control statistically for confounding variables that might influence program outcomes, the evidence for the impact of program duration is suggestive, but not conclusive.

The issue of program duration was touched upon in the Head Start Impact Study, the most recent nationwide randomized study of Head Start. A nationally representative sample of nearly 5000 newly entering three- and four-year old Head Start applicants were randomly assigned either to a treatment group that had access to Head Start services or to a control group that could receive any other non-Head Start services available in the community, chosen by their parents (U.S. DHHS, 2010). The three-year old children were eligible for two years of Head Start, while the four-year old children were eligible for one year of services. The study examined program outcomes separately for these two age cohorts. However, due to ethical concerns about possible denial of services to eligible children, one year after the randomization, children in the control group were allowed to enter the Head Start program (although typically *not* at the study sample centers), and some treatment group children left Head Start. By design, this study did not attempt to control whether three-year old children were enrolled for two years of Head Start or four-year old children were enrolled for one year (U.S. DHHS, 2010). Therefore, regretfully the study could not make a precise estimation of program duration impact. In addition, the sample statistics showed that the two age cohorts varied in their demographic characteristics, such as their ethnic distribution.

In general, the study demonstrated program impact on three-year old children's social-emotional development (e.g., less hyperactive and problem behaviors, better social skills and positive approaches to learning, and more positive relationships with parents), and these positive effects were maintained through first grade (U.S. DHHS, 2010). In contrast, four-year old children did not show these benefits. The program had minimal impact on cognitive development for both three- and four-year old children. The benefits of access to Head Start largely disappeared by first grade for the program participants as a whole. However, for three-year old children, there were a few sustained benefits (e.g., more positive relationships with parents, authoritarian parenting, and parents' less use of spanking). These results are suggestive of the influence of Head Start program duration, but the absence of randomization to different lengths of program and the demographic differences between the two age cohorts make it difficult to draw more causal conclusions.

Very few studies have made meaningful comparisons between the three- and four-year old age groups enrolled in Head Start. A careful review of the federal reports generated from the FACES studies (http://www.acf.hhs.gov/programs/opre/hs/faces/) showed that many analyses were conducted with the overall combined sample. When the program outcomes were compared and contrasted between the two age groups, usually a direct comparison was made without carefully controlling for confounding variables. In some cases, analyses were performed for each age cohort separately or just one of the cohorts (e.g., Tarullo, Aikens, Moiduddin, & West, 2010). For example, a recent secondary study of the FACES 1997 data focused on three-year old children only to examine family and program factors that predict children's academic development over time (Hindman, Skibbe, Miller, & Zimmerman, 2010). One exception is a longitudinal study of the FACES 1997 data that examines the differential academic growth trajectories between three- and four-year old children from the beginning of Head Start through first grade and the associated factors. The study found that three-year old children consistently showed more rapid growth than four-year old children, especially during kindergarten and first grade (Wen, Bulotsky-Shearer, Hahs-Vaughn, & Korfmacher, 2012).

More convincing evidence for the link between intervention duration and child outcomes comes from a federally funded public program for economically disadvantaged minority children in the Chicago Public Schools – the Chicago Child-Parent Center Program (CPC). Started in 1967, the CPC is the country's second oldest (after Head Start) federal preschool program. It offered intervention services from preschool through third grade. A quasi-experimental study of the CPC by Reynolds (1994) separated 1052 children into a control group and six different treatment combinations, allowing for testing of the impact of both intervention timing and duration on child outcomes: (a) preschool, kindergarten, and three years of primary grade intervention; (b) preschool, kindergarten, and two years of primary grade intervention; (c) preschool, kindergarten, and one year of primary grade intervention; (d) preschool and kindergarten, no primary grade follow-up intervention; (e) kindergarten and three years of primary grade intervention; and (f) kindergarten and one or two years of primary grade intervention.

The results suggested that duration of intervention exposure was significantly associated with increased reading and mathematics achievements, teacher ratings of school adjustment, grade retention, and special education placement in grades three through five, after the influence of potentially confounding variables was controlled. Full participation in the CPC program (i.e., from preschool through third grade) yielded the largest effects over time. Similar findings were repeated in a later study by Reynolds and Temple (1998). Interestingly, Reynolds (1995) also directly compared one year versus two years of preschool attendance only on the CPC children's short- and long-term outcomes, and found that two-year participants began and ended kindergarten more academically competent than one-year participants, however, the group differences disappeared in elementary grades. The quasi-experimental study of the CPC program suggests that more years of intervention is important for maintaining long-term program outcomes. Although Reynold's study (1995) showed that two years of preschool intervention can lead to better child outcomes than one year of participation, it is important to consider for how long these positive effects can be sustained.

The studies of both Abecedarian Project and CPC program involve a test of follow-up intervention in elementary schools. Given the distinct educational contexts in preschool (or infant and toddler programs) and formal education system, it is unclear if the same length of treatment in these two different settings would have similar implications in child outcomes. Researchers should be cautious in interpreting the duration impact when it pertains to different educational systems and contexts.

As mentioned, program duration is just one way of measuring intervention dosage. There are studies that focus on other aspects of intervention dosage, such as intervention intensity and parents' level of program involvement, and these constructs have been shown to predict program outcomes. For example, parents' participation with program activities (e.g., volunteering in the classroom and attending parent workshops) is associated with children's social and academic outcomes (e.g., Lamb-Parker et al., 1997; Miedel & Reynolds, 1999). Several studies with the Infant Health and Development Program (IHDP), a randomized intervention trial with low birth weight (LBW) premature infants and their parents, demonstrated that higher levels of program exposure and intensity (e.g., number of days in the child care center) were related to more favorable child and parent outcomes (Klebanov & Brooks-Gunn, 2008; Liaw, Meisels, & Brooks-Gunn, 1995; Ramey et al., 1992). However, most of these studies are correlational, and the conclusion regarding the association between dosage and program response is only suggestive. One exception is the Hill et al. study (2003) that estimated high dosage effects of the IHDP intervention by using propensity scores to match a control sample with a treatment sample that had high program participation rates (attending more than 400 days of the IHDP) on a series of demographic characteristics, and found that by age eight, the program effects on the IQ scores of children in the treatment group ranged from seven to ten points. This propensity estimation of dosage effects was substantially higher in comparison to the intention-to-treat (ITT) estimates (averaged over all participation levels).

One critical challenge with most dosage – effect studies is that they are not experimental designs (Powell, 2005). The most convincing evidence comes from experimental designs that make a direct comparison of the relative effectiveness of various levels of dosage within one intervention program, just like the design of the Chicago Parent-Child Center evaluation. The fact is that very few studies are done in this way with a large enough sample, and ethical concerns may make such an experimental study unpalatable to service communities. The IHDP study by Hill et al. (2003), with its use of propensity analysis, provides an example of an alternate pathway for examining the association between program dosage (more specifically, program intensity) and response. Our study adopted the same methodology, but focused on program duration, a different way of measuring intervention dosage.

In summary, most empirical evidence regarding the positive effects of early education intervention duration on child and family outcomes is from research of small-scale model programs, and the findings with these model programs cannot be easily generalized to large-scale public programs, such as Head Start, given program operation and implementation differences. The causal relationship between intervention duration and program effect has not been well addressed within federally funded programs. Although there have been some dosage-effect explorations with Head Start populations, including the most recent Head Start Impact Study, the review of current literature indicates that the duration of public early education interventions needs further investigation, with a large sample and a methodology that moves beyond simple correlational designs. To this end, the current study examines the nation's largest preschool intervention program – Head Start, using a nationally representative sample and a rigorous methodology, propensity score analysis, in order to draw conclusions about the impact of Head Start duration on child outcomes.

2. Detecting program duration impact using propensity score analysis

The present study aimed to address the question of whether children receiving two years of Head Start services would have better academic and social outcomes in comparison to children receiving one year of Head Start. The method of propensity score analysis matches demographically comparable samples of children with different program durations (one year vs. two years) to make a less biased estimation of duration effects. Though propensity score analysis has existed for a few decades (Rosenbaum & Rubin, 1984), it has recently gained popularity in education research due to the increased emphasis on controlling for extraneous variables in quasi-experimental studies. It provides the next best alternative when it is not possible to implement a randomized controlled study to obtain an unbiased estimate.

Specifically, propensity score analysis has several convenient features that make it a robust method for obtaining less biased estimates. For example, in order to achieve an effect estimation of the variable of interest (e.g., program duration), the propensity score model matches the comparison groups on as many covariates as possible in order to control for confounding effects that might also contribute to the outcome differences (Zanutto, Lu, & Hornik, 2005). The model does not have to be parsimonious and allows for the inclusion of all available observed covariates so that a good balance can be achieved in the sample groups (Rubin & Thomas, 1996). With a rich secondary dataset like FACES, the propensity score model can help to reduce many covariates into a uni-dimensional propensity score for sample matching. Unlike many other procedures that fail to converge when there are too many covariates, the underlying mathematics of propensity score analysis are designed such that many covariates can be accommodated. In addition, the propensity score models also weigh each covariate according to its importance (Dehejia & Wahba, 1999), and can take into account missing data by incorporating the missing value pattern into the creation of the propensity score (Haviland, Nagin, & Rosenbaum, 2007). Overall, the propensity score model is an efficient methodology for addressing our research question.

3. Research question and hypothesis

Using propensity score matching (a quasi-experimental design) to control for group demographic differences at baseline, the present study examined the impact of Head Start attendance duration on children's school performance in kindergarten. Between Head Start children who were eligible for two years of program services and those who were eligible for one year of program services, were there any significant differences in their academic and social outcomes at the end of kindergarten? Our hypothesis was that children enrolled in Head Start for two years will have better school performance (specifically PPVT, Woodcock–Johnson Reading Skills, Woodcock–Johnson Math Reasoning Skills, composite academic skills, preschool learning behavior, and social skills) by the end of kindergarten.

4. Method

4.1. Data source – FACES 2003

As part of the Head Start Program Performance Measures Initiative, the Office of Head Start launched the Family and Child Experience Survey (FACES), which is an ongoing, national, and longitudinal study of the characteristics, experiences, and outcomes of children and families served by Head Start, as well as the characteristics of the Head Start programs that serve them (U.S. DHHS, 2008). To date, three FACES cohorts have been completed (i.e., FACES 1997, 2000, and 2003). This current study utilized data from the most recent cohort – FACES 2003.

FACES 2003 recruited a nationally representative sample of 2400 newly entering three- and four-year old children and their families from 63 Head Start programs, 175 centers, and 337 classrooms. The sample was stratified by region of the country, Metropolitan Statistical Area status (urban or rural), percentage of minority enrollment (above or below 50%), auspice type (school-based or other), and percentage of non-English-speaking children in the program. The study gathered comprehensive data on child cognitive, social, emotional, and physical development, family characteristics, well-being and accomplishments, quality of Head Start classrooms, and characteristics of Head Start programs and staff. Data were collected through direct child assessment, parent interview, classroom observations, and teacher report.

The data collection phases occurred in Fall 2003 (wave one, initial assessment for all children who newly entered Head Start), Spring 2004 (wave two, assessment of children who continued with Head Start and those completing the program), Spring 2005 (wave three, assessment of Head Start children completing the program and kindergarten follow-up of children who completed Head Start in Spring 2004), and Spring 2006 (wave four, kindergarten follow-up of children who completed Head Start in Spring 2005). Overall, all three- and four-year old children were assessed at the beginning of Head Start, end of Head Start, and end of kindergarten. But three-year old children were assessed one more time in the middle of Head Start as they stayed in the program for two years. The current study focused on the child outcomes assessed at the end of kindergarten.

4.2. Participants

The current study used the longitudinal data of the FACES 2003 that included children who had kindergarten follow-up data to address the proposed research question. This sample consisted of 1778 children, of which 47% were three-year old and the rest were four-year old, and 49% were boys. Table 1 summarizes detailed child and family characteristics, disaggregated by child age cohorts. All these child and family covariates presented in the table were used in the following propensity score analysis.

Overall, prior to propensity matching, there appeared to be limited differences between participants who attended two years versus one year of Head Start programming. Among the 28 participant demographic covariates, only four demonstrated significant group differences, including mothers' age, initial PPVT scores, percentage of children who participated in Early Head Start, and the child ethnicity composite. The parents of four-year old children were slightly older; initial PPVT scores, which was age-standardized, favored four-year old children; a slightly higher percentage of three-year old children participated in Early Head Start programs; and a higher percentage of white and a lower percentage of black children were present in the four-year old group. Other than these four covariates, the two age cohorts were comparable on the remaining demographic characteristics (e.g., family structure, language spoken at home, income, and parent education) and parents in both groups had similar ratings on parenting styles, psychological wellness, and their level of involvement with their child and the Head Start program.

The three-year olds were eligible to receive two years of Head Start services, while the four-year olds were eligible for one year of program services. In the current sample, however, 6% ($n = 49$) of three-year olds stayed in the program for only one year and 7% ($n = 71$) of four-year olds stayed for two years. This crossover situation happens in reality, even in randomized controlled trials. In our analysis, we retained these crossover cases in their original age cohort. This intent-to-treat (ITT; Lachin, 2000) analysis is advocated as it more closely reflects reality in measuring the impact of pre-assigned program duration. The inference yielded by the ITT analysis is often of great policy interest because if a treatment is implemented widely as a matter of policy, imperfect treatment implementation will occur. Thus, ITT analysis gives an idea of the likely effects of the treatment-as-implemented in policy (Shadish, Cook, & Campbell, 2002).

4.3. Measures

School readiness outcomes. The current study focused on six academic and social outcomes assessed by the end of children's

Table 1
Descriptive statistics of child and family covariates used for propensity score matching.

	Continuous covariates		*F*-test	% of missing
	Two years of Head Start (three-year old children) Mean (*SD*)	One year of Head Start (four-year old children) Mean (*SD*)		
Mother's age by time of study	28.11 (6.05)	29.09 (6.37)	10.44**	4.9
Family size	4.58 (1.63)	4.67 (1.70)	1.13	3.7
Annual family income	16202.95 (12189.75)	16679.93 (12577.36)	.60	7.9
Maternal locus of control	15.23 (3.25)	15.03 (3.29)	1.57	3.9
Maternal depression (CES-D)	6.44 (6.06)	6.97 (6.52)	2.99	3.6
Frequency of spanking child	.87 (1.45)	.77 (1.40)	2.01	3.4
Parental warmth score	4.35 (.44)	4.36 (.44)	.12	3.2
Parental energy score	3.85 (.73)	3.83 (.72)	.28	3.3
Parental authoritative score	4.16 (.58)	4.21 (.59)	3.22	3.2
Parental authoritarian score	2.23 (.65)	2.19 (.66)	1.38	3.2
Frequency of reading to child	4.51 (2.43)	4.53 (2.37)	.02	3.4
Weekly/monthly activities	10.25 (3.65)	10.52 (3.57)	2.30	3.2
Involvement with program activities	6.33 (4.17)	6.09 (3.95)	1.43	7.6
How often missed Head Start	2.44 (.79)	2.48 (.82)	1.06	8.3
Initial performance: PPVT	61.37 (7.43)	70.28 (9.90)	366.47**	18.6

	Categorical covariates		Chi-square test	% of missing
	Two years of Head Start (three-year old children) Percentage	One year of Head Start (four-year old children) Percentage		
Participated in Early Head Start	8.7	6.2	4.40*	3.7
Child has identified disability	16.2	13.5	2.59	3.3
Child in good health	95.5	96.3	.58	3.2
Child ethnicity	25.1 White	32.8 White	20.14**	3.4
	35.0 Black	26.5 Black	14.20**	
	32.7 Hispanic	34.1 Hispanic	.42	
	7.2 Other	6.6 Other	.02	
Dual language learner	20.5	22.6	1.08	3.3
Parent education (highest degree of both parents)	19 no high school	24 no high school	14.75	3.4
	41.3 high school or GED	41.6 high school or GED		
	33.4 associate degree or some college	28.8 associate degree or some college		
	6.3 bachelor or higher	5.6 bachelor or higher		
Parent employment	9.1 two parents full-time	10.0 two parents full-time	9.64	5.7
	8.8 one full time, one part time	6.0 one full time, one part time		
	45.6 one working full time	44.0 one parent full time		
	10.3 one working part time	12.2 one parent part time		
	21.7 no parent working	23.5 no parent working		
Family structure	48.8 mother-father	46.9 mother-father	2.90	3.4
	46.3 mother only	47.6 mother only		
	1.6 father only	2.6 father only		
	3.3 neither mother nor father	2.9 neither mother nor father		
Language spoken at home	65 English	64.6 English	1.27	3.3
	30 Hispanic	31.4 Hispanic		
	5.1 Other language	4.1 Other language		
Supported by welfare	19	16.8	1.42	3.3
Parent in poor health	13.8	14.8	.37	3.3
Family domestic violence	9.4	9.5	.01	3.4
Parental criminality	19.6	23.0	2.85	4.0

* $p < .05$.
** $p < .01$.

kindergarten year: receptive vocabulary skills, emergent literacy skills, mathematic skills, composite academic skills, learning behaviors, and social competence. These measures represent a broad definition of school performance that goes beyond the narrow focus of academic-related skills.

Receptive vocabulary skills. The Peabody Picture Vocabulary Test (PPVT-III; Dunn & Dunn, 1997), a widely used language measure, was used to assess children's understanding of the meaning of words. Children were asked to say the number or point to the one of four pictures that best shows the meaning of a word that is said aloud by the assessor. This test is suitable for a wide range of ages from 2.5 through adulthood and has established age norms based on a national sample of 2725 children and adults. The published internal consistency reliability coefficients were reported ranging from .92 to .98, with test–retest reliability ranging from .91 to .94 (Dunn & Dunn, 1997). In the FACES sample, internal consistency reliability was .84 or greater. The one-parameter item response theory (IRT) *W* score of the PPVT test available in FACES 2003 dataset was used in the analyses. IRT models are based on a mathematical function related to the probability of answering an item correctly and account for both item difficulty and the ability of the test taker. The *W* score derived from these analyses allowed for a more precise comparison between age groups (as compared to the raw or standard score; Hambleton, Swaminathan, & Rogers, 1991).

Emergent literacy skills. The Basic Reading Skills cluster, which was a combination of the Woodcock–Johnson III Letter-Word Identification and Word Attack tasks (Woodcock & Johnson, 1989), assessed children's literacy skills. The Letter-Word Identification subtest (25 items) measured children's reading skills by naming letters and reading words aloud from a list. The Word Attack subtest (32 items) asked children to read nonsense words (e.g., plurp, fronkett) aloud to test their phonetic word attack skills. Therefore, the Basic Reading Skills cluster was an aggregate measure of sight vocabulary, phonics, and structural analysis. The reliability estimate for the cluster was .95 (McGrew & Woodcock, 2001). In the current sample, the internal consistency reliabilities for the Letter-Word Identification and Word Attack subtests were .91 and .88, respectively.

Mathematic skills. The Math Reasoning cluster, a combination of the Woodcock–Johnson III Applied Problems and Quantitative Concept tasks, assessed children's mathematic skills. The Applied Problems subtest consisted of 30 items, measuring children's skill in analyzing and solving practical problems in mathematics. In order to solve the problems, the child must recognize the procedure to be followed and then perform counting, addition, or subtraction operations (U.S. DHHS, 2008). The Qualitative Concept subtest included 19 items and involved oral questions about mathematical factual information and operations signs. Overall, the Math Reasoning cluster was an aggregate measure of problem solving, analysis, reasoning, and vocabulary. Its published reliability estimate was .95 (McGrew & Woodcock, 2001). In the current sample, the internal consistency reliability for the Applied Problems subtest and Quantitative Concept tasks were .88 and .77, respectively.

Composite academic skills. Through the Teacher–Child Report (TCR), teachers were asked to rate child's academic performance in language and literacy, science and social studies, and mathematics, in comparison to other children in the classroom. Each subject area was rated on a 5-point scale (1 = far below average, 2 = below average, 3 = average, 4 = above average, 5 = far above average).

Learning behaviors. The Preschool Learning Behavior Scale (PLBS; McDermott, Green, Francis, & Stott, 2000) was designed for classroom teachers to rate individual children on 24 items pertaining to learning-related behaviors. The behaviors included competence motivation (e.g., reluctant to tackle new activity; cries when faced with difficulty), attention/persistence (e.g., sticks to an activity), attitude toward learning (e.g., unwilling to accept help; little desire to please; aggressive when frustrated), and flexibility/strategy (e.g., carries out tasks to own ideas or relies on personal charm). Each item was rated on a 3-point scale (1 = not true, 2 = somewhat or sometimes true, 3 = very true or often true). The scale was correlated with other measures of child social skills, behavior problems, and cognitive ability; its test–retest reliability ranged from .80 to .94 and internal consistency reliability was above .70 (McDermott et al., 2000). The scale achieved a consistency reliability of .92 for the current study sample.

Social competence. Teachers were asked to rate how often the child engaged in cooperative classroom behaviors, such as following the teacher's direction, complimenting a classmate, making friends easily, waiting their turn in games/activities, and following rules when playing games. The items were from the Personal Maturity Scale (12 items were included; Alexander & Entwisle, 1988) and the Social Skills Rating System (Elliot, Gresham, Freeman, & McCloskey, 1988). The items were rated on a 3-point scale (1 = never, 2 = sometimes, 3 = very often), and a higher score indicating more frequent cooperative behavior (U.S. DHHS, 2008). The sample reliability of the teacher ratings was .88.

Child and family covariates. As presented in Table 1, a total of 28 child and family covariates were included in the propensity score analysis. These covariates covered a comprehensive list of variables identified in the early development and education literature that are associated with child development and learning (e.g., Damon & Learner, 2006; Shonkoff & Meisels, 2000). These variables encompassed: (a) child characteristics (e.g., ethnicity, health status, whether they had diagnosed disabilities, whether they were a dual language learner); (b) family demographic characteristics (e.g., parent education, employment status, family income, family size, marital status, parent age, maternal depression, welfare status, parental health status, and home language); (c) parenting styles (e.g., parental warmth) and parent involvement with child (e.g., frequency of reading, weekly and monthly activities with child); (d) child initial performance on receptive language skills (i.e., PPVT standardized test score in the fall of 2003, the beginning of Head Start); (e) child prior intervention experience (i.e., whether the child attended Early Head Start); and (f) the amount of Head Start services the child and family received (i.e., frequency of missing Head Start and parent participation with program activities), which has shown to be associated with child outcomes in previous intervention research (e.g., Hill & Craft, 2003; Mantzicopoulos, 2003; Miedel & Reynolds, 1999).

To collect data on sample characteristics, a few scale measures were used. For example, the 12-item, 4-point Center for Epidemiologic Studies-Depression Scale (CES-D; Radloff, 1977) was administered to assess mothers' depressive symptoms; the 7-item, 4-point Pearlin Mastery Scale (Pearlin, Menaghan, Lieberman, & Mullan, 1981) was used to measure mothers' sense of control over their lives and self-confidence in their ability to solve life problems (e.g., I have little control over things that happen to me). Four aspects of parenting style were assessed through parent interview: parental warmth (e.g., have warm initiate moments with child), energy (e.g., no energy to make child behave), authoritative style (e.g., control child by warning bad things), authoritarian (e.g., allow child to get angry with parent). Parent responses were rated on a 5-point scale (1 = exactly like me, 3 = somewhat like me, 5 = not at all like me).

Parent participation with Head Start activities was measured via parent interview (U.S. DHHS, 2008). Parents were asked to rate how often they participated in ten listed program activities (e.g., volunteered in classroom, attended Head Start social events, helped the field trip, observed in child's classroom, prepared foods or materials for special events, attended parent meetings/workshops), and their level of participation with each activity was rated on a 5-point scale (1 = not yet, 3 = several times, 5 = at least once a week). Parent involvement with their child measured the frequency of parents performing 11 weekly activities with the child (e.g., read stories, taught numbers, and played game together) and 11 monthly activities with the child (e.g., took child to library, art gallery, or museum, or visited a zoo or aquarium).

4.4. Analytic strategy

Propensity score analysis is a quasi-experimental design with non-equivalent groups (Shadish et al., 2002). It is a recommended procedure to use when the goal is estimation of causal effects using observational data (the type of data that are collected in situations where independent manipulation of the treatment condition by the researcher was not possible) (Rosenbaum & Rubin, 1984; Rubin, 1987). Stated formally, the propensity score $q(X)$ is the conditional probability of receiving one form of treatment over another, given observed characteristics and is modeled as follows:

$$q(X) = Pr \quad (Z = 1|X) \tag{1a}$$

where $Z=1$ for Treatment 1 units and $Z=0$ for Treatment 2 units (Rosenbaum, 1998; Rosenbaum & Rubin, 1985). In this study, the propensity score was defined as the conditional probability of being enrolled in Head Start for two years (Treatment 1) over one year (Treatment 2) given the observed covariates (i.e., the child and family covariates). In a randomized controlled trial, the probability for each child receiving two- or one-year program services is 0.5, so each child has an equal chance of being in either type of the program. However, in a non-randomized study using secondary data from the FACES, the probability of each child getting into two years or one year Head Start program varies depending on their demographic characteristics. In other words, this probability was calculated for each child (so each child receives a propensity/probability score) based on their observed covariates through a mathematical model, and it was not based on whether the child was actually in the two years or one year of Head Start.

In practice, the propensity score is usually estimated using the logistic regression of Z on X ($Z=1$ for Treatment 1 units and $Z=0$ for Treatment 2 units; X denotes the covariates; Rosenbaum, 1998). The logit model used in this study can be specified generally as:

$$q(X) = \frac{\log[1-e(x)]}{e(x)} = \alpha + \beta^{\mathrm{T}} f(x) \tag{1b}$$

where α and β are parameters to be estimated, $q(X)$ is the log of the odds ratio *against* enrollment in Head Start for two years; $f(x)$ is the specified function of the observed/measured covariates (Rosenbaum, 1998).

Estimating the propensity scores. In the logistic regression model, the independent variables were all the covariates that we would like to control for (see Table 1 for a list of covariates), and the dependent variable was the probability of receiving two years of Head Start (i.e., propensity score). As seen in Eq. (1b), the propensity score model collapsed all the observed covariates into a single, uni-dimensional 'covariate' – the propensity score, which was the predicted value from the logistic regression. The coefficients in the logistic regression naturally weighed the covariates according to their importance (and thus all covariates were not given equal importance in the matching). In addition, given the hierarchical nature of the FACES data (i.e., children nested within classrooms), the classroom IDs were taken into account in computing the propensity scores. The logistic regression was performed via the SAS program (v. 9.1). Sample matching was then performed based on this single propensity score. The computed propensity score helped to form several precise comparison groups in which the two-year and one-year Head Start children with similar propensity scores, indicating similar demographic backgrounds, were matched for the school outcome comparison.

Balance of the propensity scores was also examined to ensure that the two matched sample groups were comparable in the sense that they had similar distributions of covariates. This procedure ensured that we had included sufficient covariates in the propensity score model so that children in both treatment groups would be well matched. Covariate balance was examined according to the Rubin's criteria (2001): (1) the standardized difference in the mean propensity score between the two groups was near zero; (2) the ratio of the propensity score in the two groups was near 1, and (3) the ratio of the variances in the residuals of the (continuous) covariates after adjusting for the propensity score was near to 1. A few different sets of covariates were compared, such as where no initial PPVT test score was included, or where the data missing pattern was not taken into account. The sets of covariates displayed in Table 1 best met Rubin's criteria.

Matching children through quintiles. After creation of the propensity scores, the scores were used to match children in the two years of Head Start with children in the one year of Head Start. There are several matching algorithms but we chose one that was easier to use and fitted our analytic goal well. Specifically, the quintile method of matching was used wherein the estimated propensity scores, which ranged approximately from 0 to 1, were equally divided into five quintiles. Quintile 1 contained the lowest propensity values ranging from 0 to .20, suggesting the probability of attending Head Start for two years was 20% or less for children in this quintile; and quintile 5 contained the highest propensity values ranging from .80 to 1, suggesting the probability of attending two years of Head Start was 80% or more for children in this quintile. Children with a similar range of propensity scores were grouped into the same quintile. For example, three- and four-year old children who had propensity scores ranged from 0 to .21 were grouped into quintile 1. And those children receiving one or two years of Head Start within the *same* quintile were comparable to one another on all the observed covariates, but they were less comparable to children in other quintiles. In other words, the two-year and one-year Head Start children within the same quintile did not significantly differ from each other on all the child and family characteristics included in the propensity analysis. The five quintiles meant that there were five matched and comparable groups, and we can make more valid estimation of duration impact based on these five groups of children who differed in program durations but shared similar demographic backgrounds.

Sub-classifying into quintiles is a straightforward method of grouping children with comparable, though not exact, propensity scores. As opposed to conventional matching of one child in Treatment 1 (i.e., receiving two year of Head Start) to one child in Treatment 2 (i.e., receiving one year of Head Start), quintile matching is one special form of matching that groups two clusters of children whose propensity scores are similar within a certain (narrow) range, making them comparable. It is difficult to find exact matches of children with exactly the same propensity scores. Bias reduction is the greatest with the use of quintile matching in comparison to the one-on-one pair matching (Haviland et al., 2007; Ming & Rosenbaum, 2000). Additionally, Rosenbaum and Rubin (1984) showed that using five quintiles based on propensity scores can remove approximately 90% of the initial group differences in each of the observed covariates.

Addressing missing values in the covariates. In general, the covariates used in our propensity scoring had a minimal level of missingness in the FACES data (mostly below 5%; see Table 1). Missing values of the covariates were dealt with based on the recommendations by Haviland et al. (2007). Only five covariates with more than 5% of missing were considered, and they were included in the propensity score estimation as two variables. One variable was the observed covariate with sample mean as the imputed value for missingness, and the other variable was a separate binary variable indicating whether the cases had missingness or not. With these two variables, we took into account the missing value patterns in estimating the propensity score.

5. Results

We start by first showing the school readiness outcome differences between the two age cohorts *before* they were matched on the baseline covariates using propensity scores. Table 2 demonstrates the outcome comparison between the *original, unmatched* sample of three- and four-year old children, without controlling for any covariates. This comparison was based on an analysis of variance (ANOVA) with the whole sample. Statistics summarized in Table 2 suggests that the two age groups performed significantly differently only on the two Woodcock–Johnson subtests, with small effect sizes.

Table 3 displays the school readiness outcome comparison *after* the three- and four-year olds were matched on their baseline

Table 2
Group differences in kindergarten outcomes: before propensity matching on background covariates.

	Two years of Head Start (three-year old children) Mean (*SD*)	One year of Head Start (four-year old children) Mean (*SD*)	*F*-tests	*d*
PPVT	85.46 (8.14)	84.77 (10.15)	2.47	.075
Woodcock–Johnson: Reading Skills	414.91 (28.48)	408.15 (27.56)	25.12**	.241
Woodcock–Johnson: Math Reasoning	438.68 (16.19)	436.21 (16.56)	10.00**	.151
Composite academic skills	3.11 (.87)	3.02 (.83)	3.37	.106
Preschool learning behavior	51.69 (11.38)	51.75 (11.29)	.01	-.005
Social skills	17.88 (4.73)	17.71 (4.61)	.42	.036

** $p < .01$.

covariates using propensity scores. The statistics are presented for each of the five propensity score quintiles and each quintile included children who received one or two years of Head Start with comparable baseline characteristics. Quintile 1 included those children who had the lowest predicted probability of receiving two years of Head Start given the set of observed covariates, regardless of the fact that they actually did receive two years of Head Start, whereas Quintile 5 included the children with the highest predicted probability of attending two years of Head Start.

Regardless of the propensity to receive two years of Head Start, within-stratum contrasts indicated that attending two years of Head Start led to statistically significantly higher scores in children's PPVT, Woodcock–Johnson reading skills, and Woodcock–Johnson mathematic reasoning scores by the end of kindergarten, and this was true for children in all quintiles. The effect sizes indicated a moderate to large effect (*d* ranged from .27 to .96) suggesting from about 1/3 to nearly 1 standard deviation difference, favoring children who received two years of Head Start. This also dispelled the concern that with smaller sample sizes in each quintile, the power was too small to detect significant group differences.

For the teacher-rated composite academic skills in kindergarten, within-stratum contrasts indicated that in all but one quintile, children who received two years of Head Start had statistically significantly higher ratings. The effect for Quintile 1 indicated over 3/4 of one standard deviation difference between children in one year and two years of the program. Although the difference between the two treatment groups was not statistically significant for Quintile 5, the effect size suggested a moderate effect ($d = .37$). As for preschool learning behaviors, children in Quintile 1 (lowest predicted

Table 3
Group differences in kindergarten outcomes: after propensity matching on background covariates.

	Two years of Head Start (three-year old children) Mean (*SD*)	One year of Head Start (four-year old children) Mean (*SD*)	*F*-tests	*d*
	1st quintile (three-year old children, $n = 20$; four-year old children, $n = 241$)			
PPVT	94.84 (4.41)	92.51 (6.09)	2.81*	.44
Woodcock–Johnson: Reading Skills	435.85 (16.72)	419.38 (24.34)	8.79***	.80
Woodcock–Johnson: Math Reasoning	450.20 (12.09)	444.48 (12.76)	3.74*	.46
Composite academic skills	3.82 (.67)	3.25 (.76)	9.21***	.80
Preschool learning behavior	58.06 (7.46)	53.35 (10.70)	3.33*	.51
Social skills	19.78 (3.70)	18.14 (4.63)	2.13	.39
	2nd quintile (three-year old children, $n = 121$; four-year old children, $n = 215$)			
PPVT	87.35 (9.56)	84.03 (9.22)	9.76***	.35
Woodcock–Johnson: Reading Skills	419.41 (28.95)	406.24 (26.12)	17.42***	.48
Woodcock–Johnson: Math Reasoning	443.75 (16.49)	435.88 (14.30)	20.96***	.51
Composite academic skills	3.34 (.90)	3.11 (.75)	4.84**	.28
Preschool learning behavior	53.13 (11.90)	53.14 (10.70)	.00	−.00
Social Skills	18.50 (4.89)	18.41 (3.96)	.03	.02
	3rd quintile (three-year old children, $n = 205$; four-year old children, $n = 203$)			
PPVT	85.34 (8.92)	80.93 (9.67)	23.02***	.47
Woodcock–Johnson: Reading Skills	417.77 (26.95)	404.44 (28.20)	23.38***	.48
Woodcock–Johnson: Math Reasoning	440.42 (14.14)	433.56 (15.64)	21.53***	.46
Composite academic skills	3.16 (.76)	2.95 (.90)	4.88**	.25
Preschool learning behavior	53.27 (9.52)	49.95 (11.54)	7.60***	.31
Social skills	18.48 (4.11)	17.26 (4.65)	5.94**	.28
	4th quintile (three-year old children, $n = 219$; four-year old children, $n = 93$)			
PPVT	84.67 (7.61)	79.64 (8.49)	26.65***	.27
Woodcock–Johnson: Reading Skills	412.79 (29.66)	403.09 (25.20)	7.50***	.35
Woodcock–Johnson: Math Reasoning	437.39 (16.69)	430.53 (16.89)	10.78***	.41
Composite academic skills	3.06 (.93)	2.75 (.71)	5.70**	.38
Preschool learning behavior	52.21 (11.12)	51.43 (10.23)	.24	.07
Social skills	17.84 (4.83)	16.56 (4.27)	3.47*	.28
	5th quintile (three-year old children, $n = 155$; four-year old children, $n = 33$)			
PPVT	84.85 (6.18)	77.37 (9.13)	33.06***	.96
Woodcock–Johnson: Reading Skills	410.31 (28.34)	387.70 (35.81)	14.59***	.70
Woodcock–Johnson: Math Reasoning	434.99 (16.78)	419.47 (21.03)	20.73***	.82
Composite academic skills	2.91 (.84)	2.58 (.94)	2.74	.37
Preschool learning behavior	47.96 (12.60)	50.48 (12.91)	.75	−.20
Social skills	16.74 (5.38)	17.27 (5.57)	.18	−.10

* $p < .10$.
** $p < .05$.
*** $p < .01$.

probability of receiving two years of Head Start) and 3 (mid-range probability of receiving two years of Head Start) demonstrated that the outcome favored the group who stayed in the program for a longer time. The effect sizes were moderate (*d* ranged from .31 to.51). Similarly mixed results were evident for children's social skills in kindergarten. Within-stratum contrasts resulted in statistically significant differences between one versus two years of Head Start attendance for children in Quintiles 3 and 4 (mid-range and upper midrange predicted probabilities of receiving two years of Head Start), with a moderate effect ($d = .28$). Although the group difference in Quintile 1 was not significant, the effect size was moderate ($d = .39$).

In summary, *before* children with different lengths of program attendance were matched on their baseline characteristics, the outcome comparison yielded significant differences on only two Woodcock–Johnson subtests. However, *after* propensity score matching, the two groups of children matched on the demographic backgrounds showed significant differences on all six outcome measures (within different quintiles), with decent effect sizes.

6. Discussion

Using data from a nationally representative sample, this study examined performance differences measured by the end of kindergarten between Head Start children who attended two years of the program and those who attended for just one year. The findings convey a clear message that more, rather than fewer, years of Head Start would accrue greater program outcomes. This finding is consistent with some previous studies that support the general idea that interventions designed to target young children from disadvantaged backgrounds need to start earlier and continue for a longer period of time to ensure greater long-term positive impact (e.g., Love et al., 2007; Reynolds et al., 2004; Reynolds & Temple, 1998). Our results show that children who stayed in the Head Start program for a longer duration of time had larger gains (with moderate to large effect sizes for all quintiles) in all academic outcome measures, specifically receptive vocabulary, emergent literacy skills, and mathematic skills, compared to carefully matched children who participated for a shorter duration. Additionally, teachers' ratings of overall academic skills, learning behavior, and social competence were generally better for children who stayed in the program for two years.

The program duration, however, had smaller impacts on child social skills and learning behaviors than academic outcomes. This might be because social skills and learning behaviors were measured by teacher reports, which may be less reliable than the standardized measures adapted for cognitive outcome assessment. Head Start children's overall lower academic performance as compared to the general population might also lead to the teachers' more strict ratings on their social and learning competence. The Head Start Impact Study (U.S. DHHS, 2010) showed limited impact on teacher-reported social skills and positive learning approaches, and teachers reported that children in the Head Start group demonstrated more socially reticent behaviors such as shyness at the end of first grade. However, the Head Start Impact Study found greater program impacts based on parent-rated social skills and learning behaviors, which suggests that teacher ratings could be relatively strict. In addition, some behaviors, especially learning behaviors, might take a longer time to achieve a significant change in comparison to knowledge mastery, as it involves several aspects of development, such as the child's motivation, attitudes, personality, and social communication skills. Previous research has shown that adaptive learning behavior is critically important to achievement in preschool and beyond (Schaefer & McDermott, 1999). Similarly, the Head Start program considers learning behaviors a key component of school readiness (Hyson, 2008). More extensive longitudinal studies are needed to examine the change of children's social skills and learning approaches over time.

Although this study focused on the effect of program duration on child outcomes, the nature of the FACES data actually told a two-fold story, in that the study is not just about program duration, but also involves timing of services. The three-year old children not only got one additional year of intervention services, they also started the intervention at a younger age in comparison to the four-year old children. It is difficult to disentangle program duration and timing because a program that starts earlier, as long as it continues, also results in longer duration. In the early childhood years, children experience rapid changes in their cognitive, language, and motor skills and are highly sensitive to environmental impact (Shonkoff & Phillips, 2000). Extensive research evidence suggests that programs targeting children within their early years of life can produce positive long-term outcomes (Shonkoff & Meisels, 2000). Earlier entry into preventive educational programs that last for a reasonable period of time provides a greater opportunity to intervene prior to children's development of learning difficulties (Reynolds, 2004).

There is a recent national push to expand state-funded prekindergarten programs to enhance school-related academic skills and social-behavioral competence (Barnett, Hustedt, Robin, & Schulman, 2005; Howes et al., 2008). Statistics show that these state-funded programs mainly recruit four-year old children (Barnett et al., 2005). The current study suggests that public preschool programs should target children as early as possible and keep them in the programs for a longer period of time in order to maximize the educational benefit for these vulnerable children. This study did not impart truly new information, but with its convincing results, it provides strong policy justifications for public education program funding.

Early education intervention research should place greater emphasis on studying program design (e.g., duration, timing, content) to learn what program components are working effectively, and for which group of participants (Reynolds, 2004). Given the limited financial resources and children's significant need in this nation, knowledge about these dimensions of program design will help to refine the treatment and maximize program efficiency.

In this study, we only focused on program duration, but even for children within the same length of programming, the amount of intervention services they actually received could still vary greatly, based on, for example, the number of days the child missed the program, or the number of activities in which parents participated. Studying intervention dosage is a complex matter, and there are many research questions that need to be addressed (Zaslow et al., 2010). Although this study suggests that two years of Head Start is more effective than one year, would three years of program participation be better than two years, and would four years be better than three years? For interventions with the same length of duration, does it matter when children start the program? Between program timing and duration, which one is more important for child outcomes? Our study only examined one single variable – program duration, but early intervention programs have many different combinations of various intervention designs and conditions (e.g., timing, duration, intensity, delivery model). More robust experimental research is needed before we can have a better understanding of how to best serve children in need.

A cleaner way to examine duration impact would be randomly assigning children at the same age to different lengths of program periods in order to do a causal comparison. The design of the FACES, however, does not allow such direct comparisons. In reality, the "gold standard" of program evaluation – the randomized experiment – is difficult to carry out. Experimental research has the capacity to draw conclusions about cause and effect. However, public services, like Head Start programs, are never randomly assigned, and their effectiveness may well depend on families' uptake of services – thus introducing the possibility of selection bias. For example, three-year old children were eligible for two years of Head Start services, but some of them dropped the program earlier. How long the child stays in the program and how many services they would take is a decision made by the parents and is determined by many family factors. Therefore, research methods that take into account real-life circumstances and complement results from experimental design are needed for the field of intervention program evaluation (McCall & Green, 2004).

Propensity score matching, as adopted in this study, helps to avoid the limitation of lack of a true randomized treatment and control group comparison, and allows for a within-treatment analysis of different program duration impacts on child outcomes. This method is promising in addressing some challenges that are popular in intervention research, such as self-selection of level of participation with programs. Our study suggests that before the three- and four-year old children were matched by propensity scores, only the Woodcock–Johnson tests were significantly different between the two age cohorts. However, once children were matched through the propensity scores to become comparable on their demographic characteristics, we see substantial differences between groups.

However, there are limitations to the present study. First, this was secondary analysis of existing data and we were limited by the measures available through the FACES 2003 datasets. Second, although propensity score matching is a 'next best case' when randomization is not possible, a limitation with this procedure is that only observed covariates can be analyzed. In comparison, true randomization ensures that systematic group differences do not exist based on both observed and unobserved covariates. However, sensitivity analysis (as was conducted in this study) can be performed to determine the probability that relevant but unobserved covariates were excluded from the propensity model (Rosenbaum, 1991). In conclusion, the authors hope that this current study can shed light on both the importance of intervention design research and the exploration of rigorous research methods that well serve the field of intervention evaluation.

References

Alexander, K. L. & Entwisle, D. R. (1988). Achievement in the first two years of school: Patterns and processes. *Monographs of the Society for Research in Child Development*, *53* (2, Serial No. 218)

Barnett, W. S. & Hustedt, J. T. (2011). Improving public financing for early learning programs. *Preschool Policy Brief*, (23). Retrieved from http://nieer.org/resources/policybriefs/24.pdf

Barnett, W. S., Hustedt, J. T., Robin, K. B. & Schulman, K. L. (2005). *The state of preschool: 2004 preschool yearbook*. New Brunswick, NJ: NIEER.

Belfield, C. R., Nores, M., Barnett, S. W. & Schweinhart, L. J. (2006). The high/scope perry preschool program: Cost–benefit analysis using data from the age-40 follow-up. *Journal of Human Resources*, *41*(1), 162–190.

Berlin, L. J., O'Neal, C. R. & Brooks-Gunn, J. (1998). What makes early intervention programs works? The program, its participants, and their interaction. *Zero To Three*, *18*(4), 4–15.

Bogard, K. & Takanishi, R. (2005). PK-3: An aligned and coordinated approach to education for children 3 to 8 years old. *Social Policy Report*, *19*(3), 3–24.

Burger, K. (2010). How does early childhood care and education affect cognitive development? An international review of the effects of early interventions for children from different social backgrounds. *Early Childhood Research Quarterly*, *25*, 140–165.

Camilli, G., Vargas, S., Ryan, S. & Barnett, S. W. (2010). Meta-analysis of the effects of early education interventions on cognitive and social development. *Teachers College Record*, *112*(3), 579–620.

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*, 684–698.

Campbell, F. A. & Ramey, C. T. (1995). Cognitive and school outcomes for high risk African-American students at middle adolescence: Positive effects of early intervention. *American Educational Research Journal*, *32*, 743–772.

Damon, W., & Learner, R. M. (Eds.). (2006). *Handbook of child psychology*. New York, NY: Wiley.

Dehejia, R. H. & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, *94*, 1053–1062.

Dunn, L. M. & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test* (3rd ed.). Bloomington, MN: Pearson Assessments.

Elliot, S. N., Gresham, F. M., Freeman, R. & McCloskey, G. (1988). Teacher and observer ratings of children's social skills: Validation of the social skills rating scales. *Journal of Psychoeducational Assessment*, *6*, 152–161.

Gory, K. M. (2001). Early childhood education: A meta-analytic affirmation of the short- and long-term benefits of educational opportunity. *School Psychology Quarterly*, *16*, 9–30.

Guralnick, M. J. (1997). Second-generation research in the field of early intervention. In M. J. Guralnick (Ed.), *The effectiveness of early intervention* (6th ed., pp. 3–20). Baltimore, MD: Brookes.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Press.

Haviland, A., Nagin, D. S. & Rosenbaum, P. R. (2007). Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods*, *12*(3), 247–267.

Hill, J. L., Brooks-Gunn, J. & Waldfogel, J. (2003). Sustained effects of high participation in an early intervention for low-birth-weight premature infants. *Developmental Psychology*, *39*, 730–744.

Hill, N. E. & Craft, S. A. (2003). Parent-school involvement and school performance: Mediated pathways among socioeconomically comparable African American and Euro-American families. *Journal of Educational Psychology*, *95*, 74–83.

Hindman, A. H., Skibbe, L. E., Miller, A. & Zimmerman, M. (2010). Ecological contexts and early learning: Contributions of child, family, and classroom factors during Head Start, to literacy and mathematics growth through first grade. *Early Childhood Research Quarterly*, *25*, 235–250.

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

Hyson, M. (2008). *Enthusiastic and engaged learners: Learning behaviors in the early childhood classroom*. New York, NY: Teachers College Press.

Klebanov, P. K. & Brooks-Gunn, J. (2008). Differential exposure to early childhood education services and mother–toddler interaction. *Early Childhood Research Quarterly*, *23*, 213–232.

Korfmacher, J., Green, B., Staerkel, F., Peterson, C., Cook, G., Roggman, L., et al. (2008). Parent involvement in early childhood home visiting. *Child Youth Care Forum*, *37*, 171–196.

Lachin, J. M. (2000). Statistical considerations in the intent-to-treat principle. *Controlled Clinical Trials*, *21*(3), 167–189.

Lamb-Parker, F., Piotrkowski, C. S., Kessler-Sklar, S., Baker, A. J., Peay, L. & Clark, B. (1997). *Executive summary: Parent involvement in Head Start*. New York, NY: NCJW Center for the Child.

Liaw, F. R., Meisels, S. J. & Brooks-Gunn, J. (1995). The effects of experience of early intervention on low birth weight, premature children: The infant health and development program. *Early Childhood Research Quarterly*, *10*, 405–431.

Littell, J. H., Alexander, L. B. & Reynolds, W. W. (2001). Client participation: Central and underinvestigated elements of intervention. *Social Services Review*, *75*, 1–28.

Love, J. M., Vogel, C., Raikes, H. H., Chazan-Cohen, R., Kisker, E. E., Constantine, J., et al. (2007, March). *Impacts of Early Head Start at the end of the program (Age 3) and two years later when children were in prekindergarten*. Poster presented at the Biennial Meeting of the Society for Research in Child Development, Boston, MA.

Ludwig, J. & Phillips, D. (2007). The benefits and costs of Head Start. *Social Policy Report*, *21*(3), 3–13.

Magnuson, K. A., Ruhm, C. & Waldfogel, J. (2007). The persistence of preschool effects: Do subsequent classroom experiences matter? *Early Childhood Research Quarterly*, *22*(1), 18–38.

Mantzicopoulos, P. (2003). Flunking kindergarten after Head Start: An inquiry into the contribution of contextual and individual variables. *Journal of Educational Psychology*, *95*(2), 268–278.

McCall, R. B. & Green, B. L. (2004). Beyond the methodological gold standards of behavioral research: Considerations for practice and policy. *Social Policy Report*, *18*(2), 3–9.

McDermott, P. A., Green, L. F., Francis, J. M. & Stott, D. H. (2000). *Preschool learning behaviors scale*. Philadelphia, PA: Edumetric and Clinical Science.

McGrew, K. S. & Woodcock, R. W. (2001). *Woodcock–Johnson III technical manual*. Itasca, IL: Riverside.

Miedel, W. T. & Reynolds, A. J. (1999). Parent involvement in early intervention for disadvantaged children: Does it matter? *Journal of School Psychology*, *37*(4), 379–402.

Ming, K. & Rosenbaum, P. R. (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics*, *56*, 118–124.

Pearlin, L. I., Menaghan, E. G., Lieberman, M. A. & Mullan, J. T. (1981). The stress process. *Journal of Health and Social Behavior*, *22*, 337–356.

Powell, D. R. (2005). Searches for what works in parenting interventions. In T. Luster, & L. Okagaki (Eds.), *Parenting: Ecological perspectives* (pp. 343–373). Mahwah, NJ: Erlbaum.

Radloff, L. (1977). The CES-D Scale: A self-report depression scale for research in the general population. *Journal of Applied Psychological Measure*, *1*, 385–401.

Ramey, C. T., Bryant, D. M., Wasik, B. H., Sparling, J. J., Fendt, K. H. & LaVange, L. M. (1992). Infant Health and Development Program for low birth weight, premature infants: Program elements, family participation, and child intelligence. *Pediatrics*, *3*, 454–465.

Ramey, C. T., Campbell, F. A., Burchinal, M., Skinner, M. L., Gardner, D. M. & Ramey, S. L. (2000). Persistent effects of early childhood education on high-risk children and their mothers. *Applied Developmental Science*, *4*, 2–14.

Reynolds, A. J. (1994). Effects of a preschool plus follow-on intervention for children at risk. *Developmental Psychology*, *30*, 787–804.

Reynolds, A. J. (1995). One year of preschool intervention or two: Does it matter? *Early Childhood Research Quarterly*, *10*, 1–31.

Reynolds, A. J. (2004). Research on early childhood interventions in the confirmatory mode. *Children and Youth Services Review*, *26*, 15–38.

Reynolds, A. J., Ou, S. & Topitzes, J. D. (2004). Pathways of effects of early childhood intervention on educational attainment and delinquency: A confirmatory analysis. *Child Development*, *67*, 1119–1140.

Reynolds, A. J. & Temple, J. A. (1998). Extended early childhood intervention and school achievement: Age 13 findings from the Chicago Longitudinal Study. *Child Development*, *69*, 231–246.

Ritblatt, S. N., Brassert, S. M., Johnson, R. & Gomez, F. (2001). Are two better than one? The impact of years in Head Start on child outcomes, family environment, and reading at home. *Early Childhood Research Quarterly*, *16*, 525–537.

Rosenbaum, P. R. (1991). Discussing hidden bias in observational studies. *Annals of Internal Medicine*, *115*, 901–905.

Rosenbaum, P. R. (1998). Propensity score. In P. Armitage, & T. Colton (Eds.), *Encyclopedia of biostatistics* (pp. 3551–3555). New York, NY: Wiley.

Rosenbaum, P. R. & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*(387), 516–524.

Rosenbaum, P. R. & Rubin, D. B. (1985). Constructing a control group using multivariate matching sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33–38.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, *2*(3–4), 169–188.

Rubin, D. B. & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics*, *52*, 249–264.

Schaefer, B. A. & McDermott, P. A. (1999). Learning behavior and intelligence as explanations for children's scholastic achievement. *Journal of School Psychology*, *37*, 299–313.

Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). *Experimental and quasi-experimental design for generalized causal inference*. Boston. MA: Houghton-Mifflin.

Shonkoff, J. P. & Meisels, S. J. (2000). *Handbook of early childhood intervention* (2nd ed.). New York, NY: Cambridge University Press.

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

Steuerle, C. E., Reynolds, G. & Carasso, C. (2007). *Investing in children*. Washington, DC: The Partnership for America's Economic Success.

Tarullo, L., Aikens, N., Moiduddin, E. & West, J. (2010). *A second year in Head Start: Characteristics and outcomes of children who entered the program at age three*. Washington, DC: U.S. Department of Health and Human Services.

Temple, J. A. & Reynolds, A. J. (2007). Benefits and costs of investments in preschool education: Evidence from the Child-Parent Centers and related programs. *Economics of Education Review*, *26*(1), 126–144.

U.S. Department of Health and Human Services. (2008). *Head Start Family and Child Experiences Survey (FACES) 2003 Cohort: User's guide*. Washington, DC: Author.

U.S. Department of Health and Human Services. (2010). *Head Start Impact Study: Final report*. Washington, DC: Author.

Wen, X., Bulotsky-Shearer, R. J., Hahs-Vaughn, D. L., & Korfmacher, J. (2012). Examination of Head Start program quality: Combining classroom quality and parent involvement to understand children's vocabulary, literacy, and mathematics achievement trajectories. *Early Childhood Research Quarterly*, *27*, 640–653.

Woodcock, R. W. & Johnson, M. B. (1989). *Woodcock–Johnson Psychoeducational Battery – Revised*. Allen, TX: DLM Teaching Resources.

Zanutto, E., Lu, B. & Hornik, R. (2005). Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *Journal of Educational and Behavioral Statistics*, *30*(1), 59–73.

Zaslow, M., Anderson, R., Redd, Z., Wessel, J., Tarullo, L. & Burchinal, M. (2010). *Quality dosage, thresholds, and features in early childhood settings: A review of the literature, OPRE 2011-5*. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Zigler, E. & Styfco, S. (1994). Head Start: Criticisms in a constructive context. *American Psychologist*, *49*(2), 127–132.