A Randomized Control Trial of a Statewide Voluntary Prekindergarten Program on Children's Skills and Behaviors through Third Grade

Research Report

Mark W. Lipsey, Ph.D.
Dale C. Farran, Ph.D.
Kerry G. Hofer, Ph.D.

September, 2015
Revised September 29

PEABODY research institute

VANDERBILT PEABODY COLLEGE

Director
Mark W. Lipsey, Ph. D.

Senior Associate Director
Dale C. Farran, Ph. D.

Associate Director
Sandra Jo Wilson, Ph. D.

Peabody Research Institute

Vanderbilt University

The mission of the Peabody Research Institute is to conduct research aimed at improving the effectiveness of programs for children, youth, and families. Using field research, program evaluation, and research synthesis (meta-analysis), our faculty and staff help determine which programs are actually making a difference in the lives of the people they serve.

Recommended Citation:

Lipsey, M. W., Farran, D.C., & Hofer, K. G., (2015). *A Randomized Control Trial of the Effects of a Statewide Voluntary Prekindergarten Program on Children's Skills and Behaviors through Third Grade* (Research Report). Nashville, TN: Vanderbilt University, Peabody Research Institute.

Funding Source:

The research reported here was supported by Grant #R305E090009 from the Institute of Education Sciences, U. S. Department of Education for the study titled "Evaluating the Effectiveness of Tennessee's Voluntary Pre-Kindergarten Program."

Contact Us:
Phone: 615.322.8540
Fax: 615.322.0293

Mailing Address:
Peabody Research Institute
230 Appleton Place
PMB 181
Nashville, TN 37203-5721

Delivery Address:
Peabody Research Institute
1930 South Drive
Room 410A
Nashville, TN 37212

http://peabody.vanderbilt.edu/research/pri

Staff and Contact Information

Peabody Research Institute, Vanderbilt University:
Staff Currently Associated with the TN-VPK Evaluation Project

Principal Investigator: Mark W. Lipsey, Director, Peabody Research Institute; Mark.Lipsey@vanderbilt.edu.
Co-Principal Investigator: Dale C. Farran, Senior Associate Director, Peabody Research Institute; Antonio and Anita Gotto Chair in Teaching and Learning; Dale.Farran@vanderbilt.edu.
Research Associate: Caroline Christopher, PhD, caroline.h.christopher@vanderbilt.edu
Project Coordinator: Janie Hughart
Research Analysts: Rick Feldser, Ilknur Sekmen.
Doctoral Student. Alvin Pearman
Assessors and Observers across Tennessee.

Acknowledgements
Thanks for assistance from the Tennessee Department of Education and the individuals below without whom this study would have been impossible:

Connie Casha, Director, Office of Early Learning, Division of Curriculum and Instruction; Connie.Casha@tn.gov.
Robert Taylor, Consultant and former Superintendent of Bradley County Schools; taylor_robtl@yahoo.com.
Bobbi Lussier, Executive Director, Office of Student Teaching/Teacher Licensure, Middle Tennessee State University; former Assistant Commissioner of Special Populations.

Thanks to the Tennessee's Consortium on Research, Evaluation, and Development (http://www.tnconsortium.org/) for assistance in obtaining and interpreting data from the Tennessee education data system.

Special thanks and acknowledgement go to several individuals who have been invaluable to this project over many years: Carol Bilbrey who served as the Project Manager during the first five years of the project and whose insights and skills in coordinating with school systems were immeasurably helpful; Janie Hughart who stepped in to lead the project when Dr. Bilbrey retired; Nianbo Dong who served as a data analyst for three years of the project; and Georgine Pion whose assistance with the multiple imputations cannot be overvalued.

Table of Contents

A Randomized Control Trial of the Effects of the Tennessee Voluntary Prekindergarten Program on Children's Skills and Behaviors through Third Grade

Executive Summary

In 2009, Vanderbilt University's Peabody Research Institute, in coordination with the Tennessee Department of Education's Division of Curriculum and Instruction, initiated a rigorous, independent evaluation of the state's Voluntary Prekindergarten program (TN-VPK). TN-VPK is a full-day prekindergarten program for four-year-old children expected to enter kindergarten the following school year. The program in each participating school district must meet standards set by the State Board of Education that require each classroom to have a teacher with a license in early childhood development and education, an adult-student ratio of no less than 1:10, a maximum class size of 20, and an approved age-appropriate curriculum. TN-VPK is an optional program focused on the neediest children in the state. It uses a tiered admission process with children from low-income families who apply to the program admitted first. Any remaining seats in a given location are then allocated to otherwise at-risk children including those with disabilities and limited English proficiency.

The evaluation was funded by a grant from the U. S. Department of Education's Institute of Education Sciences (R305E090009). It was designed to determine whether the children who participate in the TN-VPK program make greater academic and behavioral gains in areas that prepare them for later schooling than comparable children who do not participate in the program. It is the first prospective randomized control trial of a scaled up state-funded, targeted pre-kindergarten program that has been undertaken.

The current report presents findings from this evaluation summarizing the longitudinal effects of TN-VPK on pre-kindergarten through third grade achievement and behavioral outcomes for an Intensive Substudy Sample of 1076 children, of which 773 were randomly assigned to attend TN-VPK classrooms and 303 were not admitted. Both groups have been followed since the beginning of the pre-k year.

Research design. There are several components to the overall research design for this evaluation. The component reported here, and the one that provides the strongest test of the effects of TN-VPK, is a randomized control trial in which children applying to TN-VPK are admitted to the program on a random basis. The TN-VPK programs participating in this part of the evaluation study were among those where more eligible children were expected to apply for the program than there were seats available. Under such circumstances, only some applicants can be admitted and, of necessity, some must be turned away. The participating programs agreed to make this decision on the basis of chance, a process rather like randomly selecting names out of a hat, to determine which children would be

admitted. This procedure treats every applicant equally and, as a result, no differences are expected on average between the characteristics of the children admitted and those not admitted. Comparing their academic and behavioral outcomes after the end of the pre-k school year, then, provides a direct indication of the effects of the TN-VPK program on the children who were admitted. Comparing their achievement and behavioral trajectories through third grade provides a test of the persistence of any pre-k effects.

To implement this procedure, TN-VPK programs across Tennessee that expected more applicants than they could accommodate and were willing to participate in the evaluation submitted lists of eligible applicants to the researchers at the Peabody Research Institute. The research team shuffled each list into a random order and the TN-VPK program staff were asked to fill the available seats by first offering admission to the child at the top of the list and then going down the list in order until all the available seats were filled. Once a program had admitted enough children to fill its seats, any remaining children were put on a waiting list and admitted, in order, if an additional seat became available. Those on the waiting list who were not admitted to TN-VPK became the control group for the study.

This procedure was used for two cohorts of children, TN-VPK applicants for the 2009-10 and 2010-11 school years, and resulted in more than 3000 randomly assigned children. Both the children who participated in TN-VPK and those who did not are being tracked through the state education database, and information on various aspects of their academic performance and status is being collected each year. State achievement test data will be available for the first time on this larger sample in late fall of 2015. In addition, parental consent was obtained for a portion of this randomized sample, referred to as the Intensive Substudy. A total of 1076 children in the Intensive Substudy were directly assessed by the research team with a battery of early learning achievement measures, and were rated by their teachers, in each year of the study.

Funding from the Institute of Education Sciences supported the research through the third grade year and the findings for this report. New funding from the National Institutes of Health will allow us to continue to track children through the 7th grade.

Outcome measures. The outcome measures used to assess the effects of TN-VPK were divided into two groups. One group consisted of measures of achievement in the areas of emergent literacy, language, and math. The second group included measures of student behavior other than academic achievement that is often referred to as non-cognitive outcomes. This second group is especially relevant for assessing the longer-term effects of TN-VPK because other longitudinal studies of early childhood education programs have found that effects on cognitive outcomes often fade after the end of the program while cumulative effects on non-cognitive outcomes emerge over time.

Measures of Cognitive Achievement Outcomes. Academic gains of the children in the Intensive Substudy sample were measured with a selection of standardized tests from the

Woodcock Johnson III Achievement Battery. These were individually administered at the beginning and end of the pre-k year, and in the spring of the kindergarten through third grade years afterwards. The tests assessed early literacy, language, and math skills and included the following:

Literacy

Letter-Word Identification: Assesses the ability to identify and pronounce alphabet letters and read words.

Spelling: Assesses prewriting skills, such as drawing lines and tracing, writing letters, and spelling orally presented words.

Language

Oral Comprehension: Assesses children's ability to fill in a missing word in a spoken sentence based on semantic and syntactic cues.

Picture Vocabulary: Assesses early language and lexical knowledge by asking the child to name objects presented in pictures and point to the picture that goes with a word.

Passage Comprehension (not used in pre-k): Assesses reading comprehension through matching picture or text representations with similar semantic properties.

Math

Applied Problems: Assesses the ability to solve small numerical and spatial problems presented verbally with accompanying pictures of objects.

Quantitative Concepts: Assesses quantitative reasoning and math knowledge by asking the child to point to or state answers to questions on number identification, sequencing, shapes, symbols, and the like.

Calculation (not used in pre-k): Assesses mathematical computation skills through the completion of visually-presented numeric math problems.

WJ Composite

The scores on the above tests were summarized in two composite measures that averaged them together to create overall measures of children's combined achievement in literacy, language, and math. One composite score combined the 6 tests given each year and the other also added the two tests given only in kindergarten and beyond.

Measures of Non-Cognitive Outcomes. In addition, reports of the children's work-related skills and behavior were obtained from their kindergarten teachers early in the fall of the school year after pre-k and from their first through third grade teachers near the end of the each grade.

Two teacher rating instruments were used for this purpose:

Cooper-Farran Behavioral Rating Scales: Teacher ratings for each child on two scales:

- *Work-Related Skills*: The ability to work independently, listen to the teacher, remember and comply with instructions, complete tasks, function within designated time periods, and otherwise engage appropriately in classroom activities.
- *Social Behavior*: Social interactions with peers including appropriate behavior while participating in group activities, play, and outdoor games; expression of feelings and

ideas; and response to others' mistakes or misfortunes.

Academic Classroom and Behavior Record: Teacher ratings for each child on three scales:

- *Readiness for Grade Level Work*: How well prepared the child is for grade level work in literacy, language, and math skills as well as social behavior.
- *Liking for School*: The child's liking or disliking for school, having fun at school, enjoying and seeming happy at school.
- *Behavior Problems*: Whether the child has shown explosive or overactive behaviors, attention problems, physical or relational aggression, social withdrawal or anxiety, motor difficulties, and the like.
- *Peer Relations*: Whether other children in the classroom like the target child and how many close friends the target child has.

Summary of results. Results from this first randomized control trial of a state funded targeted pre-k program delivered at scale are complex. We focused our research on three primary questions.

The first question concerned the effectiveness of the TN-VPK program at preparing children for kindergarten entry. At the end of pre-k, the TN-VPK children had significantly higher achievement scores on all 6 of the subtests, with the largest effects on the two literacy outcomes. The effect size on the composite achievement measure was .32. This effect is of the same magnitude as Duncan and Magnuson (2013) reported for end of treatment effects for all pre-k programs and larger than the average of programs enacted since the 1980s. At the beginning of kindergarten, the teachers rated the TN-VPK children as being better prepared for kindergarten work, as having better behaviors related to learning in the classroom and as having more positive peer relations. They did not view the children as having more behavior problems and both groups of children were rated as being highly positive about school.

The second question addressed was whether subgroups of children were differentially affected by TN-VPK attendance. We examined a number of possible moderators of the pre-k effects and found no relationships for gender, ethnicity, or age of enrollment. The moderators we did find were driven by interactions involving mothers' education and children who at age 4 did not speak English. The TN-VPK effects were the largest for children who were learning English and whose mothers had less than a high school degree. English language learners with more educated mothers had the next largest effect size. The effects for native English speakers whether or not their mothers had a high school degree were considerably lower.

The third question we addressed involved the sustainability of effects on achievement and behavior beyond kindergarten entry. Children in both groups were followed and reassessed in the spring every year with over 90% of the initial sample located tested on each wave. By the end of kindergarten, the control children had caught up to the TN-VPK children and there were no longer significant differences between them

on any achievement measures. The same result was obtained at the end of first grade using both composite achievement measures.

In second grade, however, the groups began to diverge with the TN-VPK children scoring lower than the control children on most of the measures. The differences were significant on both achievement composite measures and on the math subtests. The moderating effects of ESL status and mothers' education were no longer significant, but it is interesting to note that whether or not ESL children experienced TN-VPK, by the end of third grade, their achievement was greater than either of the native English speaking groups of children.

In terms of behavioral effects, in the spring the first grade teachers reversed the fall kindergarten teacher ratings. First grade teachers rated the TN- VPK children as less well prepared for school, having poorer work skills in the classrooms, and feeling more negative about school. It is notable that these ratings preceded the downward achievement trend we found for VPK children in second and third grades. The second and third grade teachers rated the behaviors and feelings of children in the two groups as the same; there was a marginally significant effect for positive peer relations favoring the TN-VPK children by third grade teachers.

Conclusion. The TN-VPK program saturates the state; every county has at least one classroom and all school districts except one have endorsed the program by opening new classrooms. Thus, the structural support exists in the state to continue to explore pre-k as a means for preparing children for success in school, but we need to think carefully about what the next steps should be. It is apparent that the term pre-k or even "high-quality" pre-k does not convey actionable information about what the critical elements of the program should be. Now is the time to pay careful attention to the challenge of serving the country's youngest and most vulnerable children well in the pre-k programs like TN-VPK that have been developed and promoted with their needs in mind.

A Randomized Control Trial of the Effects of the Tennessee Voluntary Prekindergarten Program on Children's Skills and Behaviors through Third Grade

Much of the rationale for the recent expansion of prekindergarten programs in many states stems from the "widely advertised success of a few model programs" (Fitzpatrick, 2008, p. 1). The lack of evidence regarding effective current practice has led to a reliance on generalizations from the benefits found in studies of smaller, intensive programs unlikely to be duplicated today (Duncan & Magnuson, 2013). Expansion beyond those model programs has taken various forms. Bartik, Gormley, and Adelstein (2011), for example, distinguished "targeted" programs, those early intensive ones like the Perry Preschool and the Abecedarian project, from "universal" programs such as ones instituted by Oklahoma, Florida and Georgia for all 4 year olds in the state. But states like Tennessee have created targeted programs serving children from families below a certain income level, but scaled up and statewide rather than small and intensive. The qualities of the model programs responsible for their effects may be hard to maintain when they are taken to scale (Baker, 2011), and it is uncertain whether scaled up programs can deliver the benefits expected of them. No well-controlled study of the long term outcomes of a widely implemented state supported pre-k program has been conducted, much less demonstrated positive effects. The study reported here offers one contribution to filling that void.

Importance of Early Experiences for Children

Poverty in the United States creates pernicious environments for the development of young children, beginning in utero. The "fetal origins hypothesis" asserts that adverse experiences in the prenatal period can program a fetus to have metabolic characteristics associated with diseases into adulthood (Currie & Rossin-Slater, 2014). Experiences of poverty before age 5, especially, appear to have both immediate and long lasting consequences for children's academic achievement and behavior (Duncan, Magnuson, & Votruba-Drzal, 2014; Duncan, Ziol-Guest, & Kalil, 2010). Summarizing recent longitudinal studies, Almond and Currie (2010) concluded that characteristics of the child measured at age 7 explain much of the variation in later educational achievement and even subsequent earnings and employment. These realizations have provided fuel to the push for intervening with poor children before school entry in an attempt to remediate these adverse effects and alter the likely lifelong trajectories of these children.

Recognition that poverty produces an early educational disadvantage that persists throughout the school years is not a new insight. The link between educational status and poverty has been acknowledged at least since the 1960s when President Johnson began the war on poverty (Farran, 2007). That recognition motivated the creation of Head Start in 1964—with its curriculum focused on school readiness skills—and eventually culminated

in Goal 1 of the National Education Goals Panel: "By the year 2000, all children in America will start school ready to learn" (NEGP archives, 2015). Thus we have had fifty years' experience creating interventions prior to formal school entry for children whose families live in poverty. Despite these efforts, analyses by Reardon (2011) have demonstrated that the achievement gap between children in poverty and higher income children has actually grown in recent years.

Model Programs

The early childhood intervention programs many now refer to as "models" were begun in the 1960s and early 70s with the Abecedarian program, the most recent of these, beginning in 1973. Many of the original model programs focused on IQ as the target outcome and realized immediate and significant effects on IQ measures (Lazar, Darlington, Murray, Royce, & Snipper, 1982). Those positive IQ effects had generally dissipated by the end of 6th grade, sometimes earlier. Effects on individually assessed achievement measures persisted in some programs for some measures, with literacy achievement the one most likely to persist (Campbell, Pungello, Miller-Johnson, Burchinal, & Ramey, 2001; Schweinhart et al., 2005). The two programs whose participants have been followed the longest and most extensively are the Perry Preschool and Abecedarian programs. It is their long term effects on school completion, employment opportunity, marriage stability, and the like that are most often cited as the justification for further investments in preschool intervention and the basis for the claim that the value of the benefits will outweigh the costs (e.g., ReadyNation, n.d. *Business Case for Early Childhood Investments).*

With the publication in *Science* of Heckman's 2006 call for investments in early childhood education for disadvantaged children, the momentum increased dramatically within states for policy makers to create pre-k programs. Heckman based his conclusions about the benefits of such investments on analyses of the Perry Preschool program and more recent studies of the Chicago Child Parent Center program (e.g., Reynolds et al., 2011). However, none of the model programs has actually been replicated in any currently implemented program. Doing so would cost much more per child than any program currently allocates—in today's dollars it would cost $20,000 per child per year to implement the Perry Preschool program, and the cost for Abecedarian would be at least $16,000- $40,000 (Minervino & Pianta, 2014). Moreover, these programs were composed of elements unlikely to be duplicated in programs implemented at statewide scale. The CPC program extended through several years in elementary school; Abecedarian began when children were 6 weeks old, continued until kindergarten and provided full day care for 50 weeks of the year. A critical question, therefore, is whether programs with weaker components and constrained budgets implemented at scale can approximate the same effects produced by these widely cited model programs.

Evaluating Scaled Up Pre-K Programs

Investigating the effects of statewide pre-k programs in a way that will produce methodologically credible results is not a simple matter. While randomized experiments are considered the best tool for determining intervention effects (Cook, 2003), deciding which children can attend public pre-k and, more to the point, which ones cannot, on the basis of the equivalent of a coin flip is not a procedure readily embraced by programs committed to serving all who qualify. Nor as a practical matter is it easy to insert that procedure into statewide application and enrollment practices even when it may be acceptable in principle. Random assignment studies of smaller scale, such as representative pilot versions of promising programs, are more feasible, but initiating a statewide program with such pilot endeavors is rare. Despite Campbell's (1969) long ago call for an experimenting society that assesses the effects of new initiatives on a small scale before leaping to full implementation, state and local school systems are not usually willing to create policies on the basis of initial smaller experiments. While the particulars of how the policy is implemented in each state vary, universal pre-k (e.g., Georgia, Florida, Oklahoma) or state wide targeted pre-k (e.g., Tennessee, Washington State) have been rolled out on the basis of little more than faith that they will benefit the participating children. Attempts to evaluate their effectiveness have thus, of necessity, been largely after the fact and have made use of a range of different research designs.

Regression discontinuity designs (RDD). The age-cutoff version of an RDD has been one of the most widely implemented designs for investigating the effects of public pre-k programs. This design can be applied without requiring changes in procedures to any program that uses an age cutoff to determine eligibility for admission. Children with birthdays on one side of the cutoff are admitted, those on the other side must wait until the following year. The outcomes of interest can then be measured after the admitted group completes pre-k and the group in waiting is just ready to begin, and compared with statistical adjustments for the age difference. An RDD can be enacted relatively quickly and on a large scale. Moreover, it has intuitive appeal—it is easy to understand how children around the cutoff, with birthdays differing by only a few days, can be substantially similar and the logic of the RDD generalizes from that simple insight.

The first pre-k evaluation to use the RDD was a study of the Tulsa, OK, program (Gormley, Gayer, Phillips, and Dawson, 2005). Subsequently there have been quite a number of RDD studies of pre-k, many led by researchers at the National Institute of Early Education Research (see Wong, Cook, Barnett, & Jung, 2008). Most recently the pre-k program in Boston has received considerable attention based on the positive results of an RDD carried out by Weiland and Yoshikawa (2013). There are important and potentially problematic methodological issues inherent in the age-cutoff RDD (Lipsey et al., 2015) but, for the purposes of this paper, the major limitation of this design is that it does not allow for longitudinal follow up of the treatment and control groups it creates. Both groups, by

definition, experience a year of pre-k, but a year apart. To assess longer-term effects of large scale pre-k programs, therefore, some other design must be used.

Matching designs. Matching designs construct a comparison group of children who did not participate in the pre-k program and then compare their outcomes with those of a group of children who did participate. If children who participated in the program at issue and those who did not participate can be matched adequately on all the variables other than the pre-k experience that influence their outcomes, matched designs have the potential to produce credible estimates of both short and long-term pre-k effects. And, because they do not require programs to change their admission or selection procedures, they can be applied to large scale programs, given adequate data for matching and outcome assessment. It is thus not surprising that evaluations of state-funded pre-k programs have most often used matching designs, although those studies differ greatly with regard to when and how the matched group is created.

The question for these designs, of course, is whether the matching is indeed adequate; the results of matched designs can be easily biased by failure to match on one or more variables that affect the outcomes independently from pre-k participation and the variables on which the groups are matched. In practice, researchers have had difficulty making adequate matches as a result of limitations in the available data and the inevitable uncertainty about what variables are essential to match on in order to avoid bias. Simply determining which children have attended the pre-k program at issue and which have not can be problematic. Attendance may be a matter of record, but absence from the records does not always mean a child had the opportunity to attend but did not. Some researchers rely on survey responses from parents, especially to determine the preschool history of children who do not appear in the pre-k program records. But parents, especially those living with the stress of poverty, often do not know what type of program their children attended – they may know the name of the program or the teacher, but not its funding source (e.g., Head Start vs. state-funded pre-k), and may not even remember that very well if asked several years past the time of enrollment.

Another problem with creating adequate matches is determining the poverty status of the children and ensuring equivalency between the groups. Almost all the studies of state funded targeted pre-k have determined poverty status by eligibility for Free and Reduced Price Lunch (FRPL), but categorical status on that index is not a very precise indicator of the nature of the economic status of the respective families. It cannot usually be obtained at the time of pre-k 4-year-old eligibility when it is not likely to be a matter of record for children who do not then enroll in pre-k. A Texas study (Andrews, Jargowsky & Kune, 2012) and a Virginia one (Huang, Invernizzi, & Drake, 2012), for example, determined FRPL status at kindergarten entry, which is generally as close to the beginning of the pre-k year as such data are available. A study of the Tennessee program, on the other hand, created matches at each grade level through third grade based on FRPL status in that school year (SRG, 2011). Such matching assumes that FRPL status is stable and that status

in some later year is what it was at the beginning of the pre-k year, which is when the matched groups must be equivalent. Both assumptions are questionable.

An exception to this pattern of matching on after-the-fact poverty data is a study of Washington State's Early Childhood Education and Assistance Program (ECEAP). The state databases in Washington are excellent and well-coordinated, allowing Bania, Kay, Aos and Pennucci (2014) to access the Basic Food Benefits database and select children whose families were eligible for SNAP when the children were 3-4 years old. Furthermore, using the unusually well-developed Washington State databases, they were able to match the two groups on neighborhood poverty rate, primary language and several other important characteristics along with the typical gender and race variables.

The critical concern for any matched design is that the matching has not equated the groups on some variable that will influence the outcome in ways will then falsely appear to be a pre-k effect, or act to offset a real pre-k effect. This can easily happen because of the limited data that may be available for matching, but may also occur simply because researchers are not aware of a relevant factor and thus do not attempt to obtain the pertinent data and use it in the matching procedure. The most likely variables of this sort are those related to parents' motivation for enrolling their children in pre-k and whatever associated supportive attitudes and behavior they have for enhancing the social and cognitive development of their children. Matched designs that draw from pools of children whose parents enrolled them in pre-k and those who did not when both had opportunity to do so inherently involve differences in these motivational and behavioral. And, it is quite plausible that parents who make an effort to enroll their children support their children's development in other ways, ways that might produce better outcomes than their less motivated counterparts even without participation in pre-k.

For studies of state programs in districts where the program is offered that then compare children who attended with those who did not (e.g., the Andrews, Jargowsky & Kune, 2012, evaluation of the Texas targeted pre-k program), this problem of unobserved but potentially relevant family differences in orientation to education and child development opportunities is difficult to avoid. Other studies (e.g., Huston, Gupta, & Shexnayder, 2012) compare children who attended pre-k with matched children in a district that did not offer the program. But that still leaves the problem of identifying those parents in the non-offering district whose motivation and interest in pre-k were such that they would have enrolled their children had the program been available.

The results of matched designs for evaluating the effects of scaled up state-funded pre-k vary. The strongest effects were found with the two matched groups Bania et al. used to evaluate the Washington program. They compared the state test scores of the 5,000 children who attended the state pre-k to the scores of 24,000 children who did not and found 3rd to 5th grade effects on math (ESs between .14 to .16) and on reading (ESs between .17 and .26). A consideration for such a small penetration study as the one Washington is the knowledge and motivation the treatment group of parents must have had to enroll their

children. Other matched designs have not shown such strong effects. Based on the best matches they were able to make, the researchers who studied the Tulsa Preschool program reported no difference on third grade outcomes for the matched groups in one cohort and only a small effect size favoring the treatment group (.18) on third grade math scores for the matched groups in a second cohort (Hill, Gormley, & Adelstein, 2015). For the Texas programs (Andrews, Jargowsky & Kune, 2012; Huston, Gupta, & Shexnayder, 2012), the Virginia one (Huang, Invernizzi, and Drake, 2012), and an earlier Tennessee study, effects were found at kindergarten entry but were mostly gone by the end of first grade and very weak if at all present by the end of 3rd grade.

Difference in difference approach (DD). DD is a design approach that, in application to state-level pre-k programs, does not focus on individuals and their participation in the pre-k program but, rather, on changes in the state or county as the pre-k program is rolled out. The before and after differences in outcomes, e.g., on state achievement test scores, associated with implementation of the pre-k program are embedded within any other differences that occurred over the respective time period, or between areas being compared, that might also have affected the outcome variables, e.g., changes or differences in population characteristics. The challenge for this design, therefore, is to isolate the difference made in the target outcomes by pre-k implementation from all the other influential factors co-occurring with it.

Many DD studies compare state baseline characteristics to the rest of the U.S. before and after the introduction of universal preschool. This approach requires a common instrument that can be compared across states. What most DD studies use are the NAEP scores, appropriate because the same instrument is used in all states, but difficult because NAEP is not collected annually. NAEP assesses children in the 4th and 8th grades biannually. Nor does NAEP assess all the children in a state; NAEP scores truly act as a barometer of the state's functioning and cannot be disaggregated by whether responders attended pre-k some 6 years previously.

The Fitzpatrick (2008) report may be the first state wide evaluation to use this design. Her focus was on the program in Georgia that grew from 14% of 4 year old participation in 1995 to 55% in 2008. She used control variables related to per capita income, the state's rate of unemployment, the percent of the population under age 24 with a high school degree, the state's school expenditures per student and other important characteristics that could have changed across the years independently of the introduction of universal pre-k. Initial analyses indicated a narrowing of the gap in average NAEP scores in Georgia compared to the rest of the U.S. after the introduction of universal pre-k. With further analyses, she concluded that "the use of appropriate control groups and methods of inference renders the estimated relationship statistically insignificant" (p. 25). That conclusion is indicative of the fundamental limitation of DD approaches to assessing the effects of scaled up public pre-k. Rather like the left out variable problem in matching designs, DD studies can be biased if influential extraneous variables are not statistically

controlled in the analysis or if the statistical control methods are not adequate to fully account for their influence.

Similar sensitivity in the results was found in the Cascio and Schanzenbach (2013) study of the Georgia and Oklahoma programs. These programs in both states increased enrollment in pre-k by 18 to 20% for children whose mothers had a high school degree or less. Comparison of NAEP scores to those of other states before and after the introduction of universal pre-k indicated that the program was associated with somewhat higher NAEP scores in 8th grade. However, when the comparison was limited to other southern states (which apparently were making general increases in NAEP scores across the time period), the DD estimates became substantially smaller. In the end, the authors could only support a conclusion of marginal statewide effects from the introduction of universal pre-k.

An ambitious master's thesis created a state year panel data set that included the percentage of 4 year olds in the states enrolled in Head Start, state-funded pre-k, and special education preschools (Rosinsky, 2014). Rosinsky compared the 2007, 2009, and 2011 NAEP 4th grade math scores to program enrollment 6 years previously. Surprisingly she found a *negative* effect on NAEP math scores from high enrollment in public programs, with the effects being carried primarily by the state funded pre-k enrollment. In follow up analyses she omitted Florida and Vermont, states that had rapidly increased pre-k enrollment. Their omission diminished the indications of a negative effect, raising the question of whether scaling up rapidly perhaps comes at the expense of quality.

The varied results from DD studies may well stem from the inherent difficulty of statistically isolating pre-k program effects from other changes and differences that span the period over which they are ramped up rather than differences in the effectiveness of the programs studied. Collectively, these studies do not provide convincing evidence of either substantial or long lasting effects from scaled up public pre-k.

Randomized control trials. RDD, matched designs, and difference in difference approaches are designs with notable practical advantages, but these come with limitations in the scope or methodological credibility of the findings. For methodological credibility, random assignment designs are widely recognized as preferable; researchers turn to alternative designs when random assignment does not appear to be possible. Prior to the study of the Tennessee program presented here, there have been only two randomized control trials of a scaled up publicly funded pre-k program, the Head Start Impact Study and the Early Head Start Impact Study. Head Start is not a state program but a national one. The U.S. Congress, in its 1998 reauthorization, mandated a study of its effectiveness. The Head Start Impact Study began in 2002 and involved 84 grantee programs and 5000 children who applied to those programs. These programs were expected to have more applicants than spaces available to accommodate them and the children were randomly selected for admission with those not selected providing the control groups (Puma et al., 2012). The Head Start participants and nonparticipants randomly assigned by this process were then followed into 3rd grade.

The children admitted to Head Start made greater gains across the preschool year than the nonparticipating children in the control condition on a variety of outcome measures. However, by the end of kindergarten the control children had caught up so that the differences between the two groups were erased. Subsequent positive effects for Head Start children were found on one achievement measure at the end of 1st grade and another measure at the end of 3rd grade. Attempts to identify differential gains associated with program quality did not prove fruitful (Peck & Bell, 2014).

A similar pattern of results was found for the Early Head Start program. A randomized design found initial effects favoring Early Head Start participation, but those were mostly gone two years later and completely gone in a grade 5 follow up (Vogel, Xue, Moiduddin, Kisker, & Carlson, 2010).

Summary

Overall, attempts to assess the effects of scaled up public pre-kindergarten intervention programs have shown decidedly mixed findings. Moreover, the overwhelming majority of those studies have used research designs with known limitations, though the respective researchers have generally been aware of those limitations and made attempts to overcome them.

RDD studies almost universally show positive effects at the end of the pre-k year, but cannot examine effects after that and thus are silent on the question of whether those effects are sustained or other longer-term effects emerge. Matched designs show some relatively weak long term effects, but those designs suffer from the inherent difficulty of matching families who do not enroll their children in pre-k on the characteristics that motivate the parents of participating children to enroll their children. This is a factor that is most likely to favor better outcomes for enrolled children, and thus any associated bias would make pre-k programs look more effective than they actually were.

The clever and appealing difference in difference approach shows effects in some states, with weaker or opposite effects in other states. Because DD approaches are investigations of regional effects and not effects specifically for children who actually participated in the program, they provide limited evidence about how those particular children are affected and the extent to which any benefits they receive are sustained. The results of these designs are also heavily dependent on adequate statistical controls for other influences on regional performance levels for children that are difficult to convincingly establish.

The prospective studies prior to this one with follow-up past the end of the pre-k year that use a random assignment design to investigate the effects of a large scale publicly funded program are the two involving Head Start and Early Head Start. Head Start, serving 3-4 year olds, is most similar to scaled-up and targeted state funded pre-k programs. While there have been criticisms of the Head Start Impact Study (e.g., Zhai, Brooks-Gunn, &

Waldfogel, 2014), its main findings have not been refuted. This scaled up national program produced early effects that faded immediately and did not return in any overall fashion through third grade. *In short, despite the promise of substantial long-term benefits of pre-k implied by the model programs of a previous era, there is not yet any credible research evidence of a contemporary publicly funded pre-k program producing such effects.*

The research study presented here uses a random assignment design analogous to the one used in the Head Start Impact Study to investigate the effects of a statewide and state funded pre-k program at the end of the pre-k year with follow up through third grade. Whether such a program can produce better results than what might be expected given the prior research summarized here is the overarching question for this study.

The Tennessee Voluntary Prekindergarten Program

The Tennessee Voluntary Prekindergarten program (TN-VPK) is a state funded prekindergarten program offered to the neediest children in Tennessee. By statute, eligibility is restricted to children who are eligible for the federal free or reduced price lunch program (FRPL), followed by such other at-risk children as those with disabilities or English Language Learners, as space allows. TN-VPK is a full-day program that operates on the same calendar as the rest of the public school system in Tennessee. The program requires a licensed teacher and aide in every classroom, a maximum of 20 children per class, and a curriculum chosen from a state-approved list. According to the quality standards used by the National Institute for Early Education Research (NIEER), the TN-VPK program is among the top state pre-k programs, meeting 9 of the 10 NIEER benchmarks (Barnett et al., 2014). A current annual investment of nearly $90 million supports 935 TN-VPK classrooms in 135 of the 136 Tennessee school systems across all 95 counties in Tennessee. Of the 935 classrooms funded through VPK, 62 across 43 different sites are not located in public schools (6.6%). Two sites are affiliated with an Institute of Higher Education, 7 sites are affiliated with Head Start and the remaining 34 are nonprofit or for profit child care programs. All funds flow through Local Education Agencies. From its pilot year in 2004, the program has grown from serving 3,000 children to more than 18,000 as of fiscal year 2014. Despite that growth, the program enrolls fewer than half of the eligible children in the state (Grehan et al., 2011) and many school systems in the state receive more eligible applicants than they can accommodate.

In 2009 the Peabody Research Institute at Vanderbilt University launched a study of the TN-VPK program in coordination with the Division of School Readiness and Early Learning at the Tennessee Department of Education. That study has multiple components; this report describes the findings of one of those components that investigated the following research questions:

1. Does participation in TN-VPK improve the school readiness at kindergarten entry of the economically disadvantaged children served?

2. Does TN-VPK have differential effects for different subgroups of children and, if so, what are the characteristics of the children who show larger or smaller effects of TN-VPK participation?
3. Are the effects of TN-VPK participation sustained through the kindergarten, first, second, and third grade years?

Method

Procedures

This study is part of a larger TN-VPK evaluation that is comprised of two main components: a randomized control trial (RCT) and a regression discontinuity design. The RCT, in turn, consists of two overlapping parts. The *full sample* of participants in the RCT involves more than 3,000 children randomly assigned to receive an offer of admission to TN-VPK or not. These children are being followed in the state's education database with attention to such outcomes as attendance, retention in grade, disciplinary actions, and state achievement test scores. Information on their first state achievement test is not yet available, but results will be reported when it is. With parental consent, a subset of the children in that full sample, referred to as the *intensive substudy sample*, was individually assessed by the research team and rated by their teachers annually through their third grade year. The present report describes the findings for that intensive substudy. Prior research reports have more fully described the components of the overall study and presented findings from earlier waves of data collection (Lipsey et al., 2011, 2013a, 2013b).

***Random assignment*.** Many TN-VPK sites across the state have more eligible applicants than available seats thus creating a situation in which some applicants of necessity must be denied admission. For school year 2009-10, and again in 2010-11, the personnel in a number of those sites agreed to randomly select the applicants to whom they would offer admission rather than use the customary first-come first-served procedure. These programs sent their applicant lists to the researcher team where they were sorted into random order using a random number table and promptly returned. The school staff at each site was then instructed to fill their available TN-VPK seats in the order that children appeared on the randomized list. To do so, they were asked to attempt to contact a child's parents at least three times on different days of the week and at different times of the day to offer admission and determine if the parent wished to accept that offer for their child. If they were unable to contact the parent after these attempts or the parent declined the offer, staff were asked to move on to the next child on the list whose parents had not yet been contacted. Once all the slots in a given program were filled, the remaining children on the list who were not offered admission were identified as the waiting list. If a child offered admission did not show up for the program when school started, the next child in order on the randomized list was offered that place. Any children not offered admission after that point became the control group of TN-VPK nonparticipants.

Intensive substudy sample. Attempts were made to contact the parents of the children on eligible randomized lists at the beginning of the school year to request consent for periodic individual assessments of their children. Though very few parents explicitly refused consent, making contact and obtaining a response from the parents proved challenging. For the 2009-10 cohort of children on eligible randomized lists, practical constraints required that the parents be contacted through a mailing sent centrally from the Department of Education. For that cohort, the overall consent rate was 42% (46% for TN-VPK participants; 32% for nonparticipants). Because of this modest and differential consent rate, the 2010-11 cohort was added to the study and arrangements were negotiated to allow many of the parents to be approached about consent as an adjunct to the TN-VPK application process. The consent rate for this second cohort was higher with less differential between the participant and nonparticipant groups: 71% overall with 74% for TN-VPK participants and 68% for nonparticipants.

These procedures resulted in a total of 1076 children from the full randomized sample whose parents consented to their participation in the intensive substudy and for whom data were available on at least one outcome measure at the end of the pre-k year. Those children were represented on 76 randomized applicant lists created at 58 different schools in 21 districts spread widely across the state. Nineteen of the schools were near cities (10 large cities, 7 mid-size, and 2 small), 11 were in suburbs, 12 were in towns, and 16 were considered rural.

Basic descriptive data for the full randomization sample was available from the State Education Information System that allowed the characteristics of the children in the intensive substudy sample to be compared with those of the children who were not in that sample and with the full randomization sample. Table 1 reports those comparisons and shows that the consented children in the intensive substudy sample were generally representative of those in the full randomization sample on these characteristics. The children with parental consent to participate in the data collection for the intensive substudy, however, did include somewhat more white children and somewhat fewer Black and Hispanic children than the remainder of the full randomization sample, as well as somewhat fewer males, children for whom English was a second language, and children born outside the U.S. It should be noted that, because of the direct data collection from the children and parents in the intensive substudy sample, more accurate information was obtained for those children in some cases than what appeared in the State database. That more accurate data is used and reported where appropriate when only the intensive substudy sample is being considered in the analysis.

Additional information available for the children in the intensive substudy sample who participated in TN-VPK (the treatment group) shows that they attended the TN-VPK pre-k classes an average of 147 days (SD=23.8) during the pre-k year. For the children in the intensive substudy sample who did not participate in TN-VPK (the control group), information from interviews with their parents identified the alternative arrangements

that had been made for them during the pre-k year. A majority of these TN-VPK nonparticipants did not attend any center-based preschool program during the pre-k year when they were not admitted to the TN-VPK program. A little more than 59% were cared for at home, 11.5% attended Head Start, 15.1% were in private childcare, and the child care arrangements for the remainder were mixed or unknown.

Table 1: Characteristics of the Children in the Intensive Substudy Sample Compared with those not in that Sample and the Full Randomization Sample

Variable	Children in the intensive substudy sample (N=1076[a])	Children not in the intensive substudy sample (N=1949[b])	All the children in the full randomization sample (N=3025[c])
Mean age (months)	51.8	52.0	52.0
Gender (% male)	47.6	50.6	49.6
Race/ethnicity % White	55.9	46.1	49.6
Race/ethnicity % Black	22.7	27.3	25.6
Race/ethnicity % Hispanic	19.2	24.3	22.5
Race/ethnicity % Other	2.2	2.4	2.4
% English as second language	21.0	27.0	24.9
% Born outside the US	8.8	11.0	10.2

(a) Varied from 1072 to 1076 because of missing data on some variables.
(b) Varied from 1941 to 1949 because of missing data on some variables.
(c) Varied from 3013 to 3025 because of missing data on some variables.

Data collection. Children in the intensive substudy sample were individually assessed by trained assessors in the fall and spring of their pre-k year, and in the spring of each subsequent year through the end of the third grade year. Children who were not located in a public school were assessed when possible at a location convenient for the parents, including Head Start centers, libraries, parks, homes, and the like. Early in the kindergarten year and in the spring of the first, second, and third grade years, children's classroom behavior was also rated by their teachers. The ratings by the kindergarten teachers near the beginning of the kindergarten year are being treated as a pre-k outcome that reflects the school readiness of the children upon entry into formal schooling.
At least 92% of the intensive substudy sample was located and assessed in each of the four years following the intervention year, and the modest amount of attrition that did occur was very similar for TN-VPK participants and nonparticipants. Table 2 shows the number and proportion of children who received direct assessments each year.

Measures

Parent questionnaire. During the pre-k year, parents of consented children were interviewed via telephone. The interview protocol developed for the purposes of this study

was administered by project staff and asked parents about their child's preschool arrangements or alternate arrangements if their child was not in TN-VPK, their own education and employment and that of their spouse/partner when applicable, and a set of questions about the home language and literacy environment.

Table 2: Sample Retention for Each Data Collection Wave by Condition

	Year 1 (Pre-K)	Year 2 (K)	Year 3 (1st)	Year 4 (2nd)	Year 5 (3rd)
Nonparticipants	303	297 (.98)	291 (.96)	290 (.96)	280 (.92)
TN-VPK Participants	773	749 (.97)	738 (.95)	726 (.94)	714 (.92)
All Participants	1076	1046 (.97)	1029 (.96)	1016 (.94)	994 (.92)

Note: The proportion retained is shown in parentheses.

Direct assessments. Children's academic achievement was assessed with a selection of scales from the Woodcock Johnson III Achievement Battery (*WJ*; Woodcock, McGrew, and Mather, 2001) that are widely used in longitudinal research. The scales administered at the beginning and end of the pre-k year included two measures of early literacy (Letter-Word Identification and Spelling), two measures of language (Oral Comprehension and Picture Vocabulary), and two measures of early math skills (Applied Problems and Quantitative Concepts). At the end of the kindergarten year, and each subsequent year through the end of the third grade year, two additional subtests were added to the battery: another language measure (Passage Comprehension) and another math measure (Calculation).

Letter-Word Identification measured children's ability to identify and pronounce alphabet letters and read words by sight. The *Spelling* subtest measured children's ability to draw simple shapes and write orally-presented letters and words. *Oral Comprehension* measured children's ability to listen to and provide a missing key word to an orally presented passage. *Picture Vocabulary* tested children's expressive vocabulary. *Applied Problems* measured children's ability to solve numerical and spatial problems accompanied by pictures. *Quantitative Concepts* measured children's understanding of number identification, sequencing, shapes, and symbols and, in a separate section, to manipulate the number line. *Passage Comprehension* (not used in pre-k) assessed reading comprehension through matching picture or text representations with similar semantic properties. *Math Calculation* (not used in pre-k) assessed mathematical computation skills through the completion of visually-presented numeric math problems.

The longitudinally scaled W-scores from these measures were used in all analyses, though standard scores are also presented in some cases to facilitate interpretation. These various *WJ* scales were moderately to highly correlated with each other. To provide summary achievement indices, composite scores were created as the simple mean across

the individual scales. One composite score combined the original six subscales administered annually from the beginning of the pre-k year (WJ Composite6). The other combined those six with the additional two that were first administered at the end of the kindergarten year (Composite8).

Teacher ratings. Two teacher rating instruments were used by kindergarten, first, second, and third grade teachers. One measure, the Cooper-Farran Behavioral Rating Scales (Cooper & Farran, 1991), required teachers to rate each child's work-related skills and social behavior. Extensive development work has been done to validate this instrument and establish its reliability; details are reported in the CFBRS manual (Cooper & Farran, 1991). *Work-Related Skills* assessed the ability to work independently, listen to the teacher, remember and comply with instructions, complete tasks, function within designated time periods, and otherwise engage appropriately in classroom activities. The *Social Behavior* subscale assessed social interactions with peers including appropriate behavior while participating in group activities, play, and outdoor games; expression of feelings and ideas; and response to others' mistakes or misfortunes.

The second measure, the Academic Classroom and Behavior Record (ACBR; Farran, Bilbrey, & Lipsey, 2003), included teacher ratings on four scales. *Readiness for Grade Level Work* asked how well prepared the child was for grade level work in literacy, language, and math skills as well as social behavior. *Liking for School* included items about the child's liking or disliking for school, having fun at school, enjoying and engaging in classroom activities, and seeming happy at school. *Behavior Problems* indicated whether the child has shown explosive or overactive behaviors, attention problems, physical or relational aggression, social withdrawal or anxiety, motor difficulties, and the like. On the *Peer Relations* items, teachers rated whether other children in the classroom like the target child and how many close friends the child has.

Analysis

Missing data. There were at least some missing values for most of the variables of interest for the analysis. The average missing value rate across all these variables for the TN-VPK participants was 6.2% (ranging from 0.0% to 14.5%) and for the nonparticipants was 6.4% (ranging from 0.0% to 17.2%). To retain the full sample in all analyses, multiple imputation of the missing values was done separately for the participant and nonparticipant data using SAS®[1] and Mistler's (2013) procedures for multilevel multiple imputation. To facilitate convergence of the imputation models, the variables were divided within each condition into three groups run separately with the results then combined to reassemble a full data file. In each case, the missing values were imputed using a 2-level structure with children nested within their school-level randomized lists. Fifty imputed files were produced and stacked for analysis of each with the results pooled so as to include

[1] SAS is a registered trademark of SAS Institute Inc.

the uncertainty associated with the imputations in the standard error estimates used for statistical inference. These imputations produced a small number of missing value estimates that were outliers relative to the distribution of observed values. For continuous variables, imputed values falling outside Tukey's (1977) outer fence (plus or minus 3 times the interquartile range subtracted from Quartile 1 and added to Quartile 3) for the observed values were recoded to the respective outer fence. For integer values (e.g., ratings on a 7-point scale), imputed values falling outside the range from one scale step below the lowest observed value to one scale step above were recoded to those values. For a small number of dichotomous variables to be used as moderators in interaction terms in the analysis (e.g., gender), any imputed values were rounded to the nearest observed value.

Comparison conditions. The intensive subsample, which required parental consent, includes only a portion (36%) of the children in the full sample of children whose applications to TN-VPK were randomized. While there is a chance component in the division of the subsample into TN-VPK participants and nonparticipants that is advantageous for causal inference, there is also potential for selection bias stemming from factors that may have differentially influenced attainment of parental consent for each comparison group in each cohort. Another implication is that an intent to treat comparison is not possible—outcome data are missing for children who were randomized but for whom parental consent for intensive substudy data collection was not obtained.

The analysis approach taken here, therefore, is a comparison of children who participated in TN-VPK with those who did not participate irrespective of where their names fell on the respective original randomized lists. In that regard, it is an analysis of the effects of treatment on the treated (TOT) rather than an intent-to-treat analysis. In particular, children who attended a TN-VPK program for 20 days or more during the school year were designated as participants and compared with children designated as nonparticipants who attended fewer than 20 days (20 days is a TN DOE Office of Early Learning enrollment standard). Moreover, because of the potential for selection bias, this comparison was analyzed as a quasi-experiment, recognizing the importance of assessing baseline equivalence and taking what steps are possible to reduce the potential for selection bias to influence the results.

Baseline equivalence. The baseline variables for the intensive substudy sample are shown in Table 3, some of which are differentiated in ways that overlap with others (e.g., Hispanic race/ethnicity is further broken down it to subgroups for native or nonnative English speakers). As noted earlier, the consent rates were different for the first and second cohorts of children in the intensive subsample. These baseline variables were thus first analyzed to determine if they showed any differences between the cohorts. Three-level multilevel analysis was used with children nested within randomized applicant lists and lists nested within school districts. Of the 22 variables on which the cohorts were compared, the means for the two cohorts were significantly different only for number of working parents, with a mean of 1.1 for the 2009-10 cohort and 1.3 for the 2010-11 cohort.

Given this substantial baseline similarity between the cohorts, the data were combined for all subsequent analyses.

Table 3: Comparison of Participant and Nonparticipant Groups on Baseline Measures

Variable	TN-VPK participants [N=773] Mean (SD)	TN-VPK non-participants [N=303] Mean (SD)	p-value	Effect size	PS p-value[a]	PS adj. ES[b]
Age (years)	4.4 (.28)	4.4 (.29)	.533	-.04	.937	.01
Gender (1=male)	.47 (.50)	.48 (.50)	.932	-.01	.994	,00
Race/ethnicity Black (1=yes)	.21 (.42)	.19 (.43)	.449	.05	.802	-.02
Race/ethnicity Hispanic (1=yes)	.14 (.37)	.15 (.44)	.694	-.03	.303	.08
Native language English (1=yes)	.86 (.37)	.84 (.46)	.571	.04	.461	-.06
Not Hispanic, native English (1=yes)	.83 (.40)	.81 (.47)	.619	.03	.279	-.09
Hispanic, native English (1=yes)	.03 (.17)	.03 (.19)	.721	-.02	.443	.07
Hispanic, not native English (1=yes)	.11(.34)	.13 (.42)	.639	-.03	.502	.05
Not Hispanic, not native English (1=yes)	.03 (.18)	.04 (.26)	.510	-.05	.849	.02
Library card use (0-2)	.96 (.82)	.89 (.84)	.216	.09	.876	.01
Newspaper subscriptions (0-3)	.38 (.76)	.33 (.75)	.417	.06	.702	.04
Magazine subscriptions (0-2)	.29 (.50)	.26 (.51)	.423	.06	.332	-.09
Home literacy index	.16 (2.03)	-.02 (1.96)	.223	.09	.826	-.02
Mother's education (1-4)	2.16 (.72)	2.02 (.74)	**.010**	.19	.610	-.04
Number of working parents	1.25 (.62)	1.23 (.62)	.641	.03	.990	.00
WJ Letter-Word Identification	319.2 (27.0)	315.1 (27.2)	**.035**	.15	.815	-.02
WJ Spelling	350.6 (28.4)	349.3 (28.5)	.534	.04	.880	.01
WJ Oral Comprehension	444.4 (15.6)	442.9 (17.5)	.206	.09	.477	-.06
WJ Picture Vocabulary	457.1 (21.0)	454.4 (27.8	.088	.12	.329	-.08
WJ Applied Problems	392.1 (26.9)	391.6 (29.9)	.818	.02	.344	-.08
WJ Quantitative Concepts	407.6 (13.9)	407.3 (14.3)	.789	.02	.930	.01
WJ Composite6	395.2 (17.7)	393.6 (19.1)	.202	.09	.561	-.05

Notes: Age on Sept. 1 of pre-k year; Library card use (0=no card/used almost never, 1=used once or twice a year or every few months, 2=used more than once a year or at least weekly); Newspaper subscriptions (0=0, 1=1, 2=2-3, 3=>3); Magazine subscriptions (0=0, 1=1-3, 2=>3); Home literacy index = sum of the z-scores for Library card, Newspaper subscriptions, and Magazine subscriptions; Mother's education (1=less than high school, 2=high school diploma/GED, 3=associate's degree, 4=more than associate's degree); WJ= W-scores on the indicated Woodcock Johnson pretests.

(a) p-value for difference between means for participants and nonparticipants with the propensity score as a covariate.

(b) Effect size for the difference between means for participants and nonparticipants with the propensity score as a covariate.

The results of an analogous analysis on the combined cohorts for the baseline differences between the children in the TN-VPK participant and nonparticipant groups are shown in Table 3. These results demonstrated that these groups were substantially similar at baseline, but there were some notable differences that must be addressed. The groups were significantly different on the WJ Letter-Word Identification scale and on mother's education, both favoring the treatment group. There was also a difference on the WJ Picture Vocabulary scale at $p<.10$. The effect sizes indexing the magnitude of the various baseline differences, nonetheless, were relatively modest—none was greater than .19. These fall well under the Imbens and Rubin (2015) rule of thumb for baseline differences too large to adjust with covariates in a regression model (p. 277).

There was another more problematic difference between the TN-VPK participant and nonparticipant groups, however. The practicalities of arranging individual assessments for so many children under field conditions made it difficult to obtain every assessment within tight windows of time at the beginning and end of the school years, as appropriate. This was especially the case for the nonparticipants during the initial year when they were not in TN-VPK classrooms so that ad hoc arrangements had to be made with the parents to meet and assess them at some agreed location. As a result, the timing of assessments was variable and, in particular, it was not possible to obtain baseline pretest assessments as early in the school year as desired. Table 4 shows the mean days from the date on which the respective TN-VPK classes began to the date on which each wave of assessments was administered. There is considerable variability in the timing, as indicated by the standard deviations, and an unfortunately long average lag before it was possible to obtain pretest assessments for the children in both groups. Most notably, there are significant differences between the participant and nonparticipant groups in the timing, especially during the early waves, that is represented by large effect sizes.

Table 4: Comparison of Participant and Nonparticipant Groups on Timing Variables

Time from School Start Date	TN-VPK participants [N=773] Mean (SD)	TN-VPK non-participants [N=303] Mean (SD)	p-value	Effect Size	PS p-value[a]	PS adj. ES[b]
Days to pretest	71 (22.8)	86 (30.8)	**.000**	-0.61	.607	.03
Days to pre-k posttest	267 (13.5)	279 (20.2)	**.000**	-0.79	.604	.03
Days to K follow-up	626 (21.4)	629 (22.2)	.111	-0.11	.243	.02
Days to 1st grade follow-up	987 (26.4)	990 (29.0)	.110	-0.11	.780	.02
Days to 2nd grade follow-up	1335 (26.5)	1337 (30.0)	.256	-0.08	.505	.05
Days to 3rd grade follow-up	1695 (28.7)	1696 (43.5)	.910	-0.01	.948	.01

(a) p-value for difference between means for participants and nonparticipants with the propensity score as a covariate.
(b) Effect size for the difference between means for participants and nonparticipants with the propensity score as a covariate.

In consideration of these differences, and the few lesser ones found for the child and family characteristics that are shown in Table 3, we constructed propensity scores to assist with the task of statistically matching the groups and reducing any bias in the effect estimates that might be caused by these initial differences.

Propensity scores. The propensity scores were created via a multilevel logistic regression predicting treatment condition with children nested in their school-level randomized lists and those nested within district. The selection of predictor variables for that model focused especially on the timing variables shown in Table 4, all of which were included. Moreover, because the rate of change may have been different for the TN-VPK participants and nonparticipants during the lag time prior to pretest, an interaction term was included for that lag time crossed with baseline scores on the WJ Composite achievement measure, which was itself also included as a separate predictor. Also included was a selection of the descriptive variables for children and families shown in Table 3 (age, gender, race/ethnic subgroup, the home literacy index, mother's education, and number of working parents). In recognition of the varying consent rates across the randomized lists and the two cohorts, Level 2 variables were added for cohort and the participation rates for the treatment and control groups at each school, as well as the interaction between them.

The propensity scores were created as a predicted probability of being in the TN-VPK participant group for each child. The overlapped completely between the participant and nonparticipant groups, providing a broad range of common support, and required no trimming. Those scores showed linear relationships with the composite achievement measures across the longitudinal waves and we elected to use them as a covariate in all the statistical analyses estimating intervention effects. A check on the extent to which these propensity scores used in this manner reduced the baseline differences of concern was made by re-estimating baseline differences by condition with the propensity scores as the sole covariate in the regression models. The last two columns of Table 3 and 4 show the *p*-values and effect sizes that resulted with the propensity score adjustment. As can be seen there, this procedure was quite effective. With the propensity score covariate in the model, there were no statistically significant differences on any of the baseline variables and the corresponding propensity score adjusted effect sizes were quite small with none exceeding .10 and most well below that.

Results

TN-VPK Effects at the End of the Pre-K Year

The first research question this study was designed to address is whether TN-VPK improved the school readiness of the participating children over the course of the pre-k year. The indicators of school readiness chosen for this purpose were the Woodcock Johnson achievement measures of early literacy, language, and math skills described earlier. In addition, we asked kindergarten teachers to rate the children near the beginning

of the kindergarten year on a battery of scales asking about the children's work-related and social behavior, their feelings about school, and how well prepared to participate in kindergarten the teacher thought the child was.

The Woodcock Johnson achievement measures yield three kinds of scores—raw scores, normed standard scores, and longitudinally scaled W-scores. The statistical analysis of TN_VPK effects was conducted with the W-scores across all the waves of measurement. The standard score, however, is the more familiar form and may be easier for many readers to interpret. For descriptive purposes, therefore, standard score values are also shown for some analyses.

Using the WJ W-scores, the TN-VPK effects on the achievement measures at the end of the pre-k year were estimated in a three-level model with children nested in their school-level randomizations and schools nested in districts. The propensity scores were used as a covariate along with the pretest of the respective outcome measure and a selection of baseline child and family characteristics. The latter were included to allow for the use of consistent analytic models for moderator analysis involving any of those characteristics as well as to supplement the propensity scores as a means to ensure as much baseline equivalence as possible. Additionally, including pretests as covariates, with their relatively large correlations with the posttests, enhanced the statistical power of the analyses.

Table 5 shows the full analysis results for the WJ Composite6 outcome, which characterizes the overall pattern of achievement effects.[2] That analysis showed a statistically significant positive effect of TN-VPK on this overall average of the six individual scales used as achievement outcomes. Table 6 provides additional detail about this finding and summarizes the results of analogous analyses for each of the individual WJ scales. As indicated there, the effects on all the measures except Oral Comprehension were statistically significant at the .05 level and the p-value for Oral Comprehension fell under .10. Table 6 also shows the standardized mean difference effect sizes that correspond to the regression coefficients that estimate the difference between the posttest means for the TN-VPK participants and nonparticipants in W-score units.

Standardized effect sizes are one way to characterize the magnitude of the effects represented by the effect estimates the regression analysis yields. However, they compare the groups on the posttest only and, as such, provide no indication of the nature of the relative performance improvements by each group over the course of the pre-k year. Table 6, therefore, presents a variant on the effect size picture that is somewhat more informative. The covariate-adjusted pretest and posttest means for each group were

[2] The results presented here and in the sections below on effects also reported earlier in technical reports (Lipsey et al., 2011, 2013a, 2013b) are somewhat different than in those earlier reports, though their pattern is much the same. These differences stem from improvements in the imputation procedure and refinements in the propensity scores and other aspects of the analytic models aimed at better controlling the influence of baseline differences, especially regarding timing of measurement, as discussed above.

extracted from the analysis as one way to describe change. These involve the same covariates, other than the pretest itself, and thus are comparable. By standardizing those pre-post mean differences with the same pooled posttest standard deviation used for the more conventional effect size index, differential growth as well as the posttest differences it produces can be depicted.

Table 5: Full Analysis Results for the WJ Composite6 Outcome Measure at the End of the Pre-k Year

	Coefficient	Standard error	t-value	p-value
Intercept	91.7	7.14	12.86	.000
Propensity score	5.92	1.46	4.06	.000
Composite6 pretest	.791	.018	43.65	.000
Age (years)	-.836	.946	-.88	.377
Gender (1=male)	-.177	.522	-.34	.734
Race/ethnicity Black	1.15	.696	1.65	.100
Hispanic, native English	1.22	1.52	.80	.423
Hispanic, not native English	2.59	.933	2.78	.005
Not Hispanic, not native English	.289	1.38	.21	.834
Home literacy index	.049	.138	.35	.723
Mother's education	.422	.389	1.09	.278
Number of working parents	.065	.418	.16	.876
TN-VPK participation	**5.32**	**.753**	**7.06**	**.000**

Notes: Age on Sept. 1 of prek year; Home literacy index = sum of the z-scores for Library card, Newspaper subscriptions, and Magazine subscriptions; Mother's education (1=less than high school, 2=high school diploma/GED, 3=associate's degree, 4=more than associate's degree).

The last three columns of Table 6 show this effect size variant. They reveal, first, that both groups of children showed performance improvements during the pre-k year, though the amount in this standard deviation metric varied for the different achievement measures. The pre-post gains on the language measures, for instance, were smaller than those on the literacy and math measures. Relative to the gains made by the nonparticipants, those made by the TN-VPK participants were proportionately much greater on most of these measures, with increases ranging from 20% to 83%. However, one of the largest proportionate gains was made on a performance measure that did not improve very much for either group—Picture Vocabulary.

Table 6: TN-VPK Effect Estimates for Pre-K Gain on Woodcock Johnson Achievement Measures

Outcome	TN-VPK effect estimate in W-score units	p-value	Effect size	Effect size for non-participant gain	Effect size for TN-VPK participant gain	% Increase in Gain for TN-VPK participants
WJ Composite6	5.32	<.001	.32	.74	1.06	44%
Literacy Measures						
Letter-Word Identification	10.77	<.001	.41	.60	1.01	68%
Spelling	7.22	<.001	.29	.80	1.09	36%
Language Measures						
Oral Comprehension	1.50	.093	.09	.44	.53	20%
Picture Vocabulary	3.66	<.001	.20	.24	.44	83%
Math Measures						
Applied Problems	4.03	.005	.17	.61	.78	28%
Quantitative Concepts	4.32	<.001	.27	.68	.96	40%

Another way to characterize the nature of these findings on achievement measures is to compare them with the results of other studies of pre-k effects. Summarizing the immediate effects of 84 pre-k programs, Duncan and Magnuson (2013) estimated the simple average effect size at the end of the pre-k year as .35. However, that includes earlier studies going back to the 1960s. Programs researched since the 1980s had an average effect size of .16

Teacher ratings. Kindergarten teachers in classrooms that included children from the intensive substudy sample were asked to rate those children near the beginning of the kindergarten year on the rating scales described earlier that focused on their behavior in the classroom and the teacher's perception of how prepared they were for kindergarten, i.e., their school readiness. No information was provided to the teachers about which of those children had participated in TN-VPK and which had not. The timing for these ratings was aimed at a period a few weeks past the start of the school year, lagged enough so the teachers would have a chance to become familiar with the children but not so much that the kindergarten experience itself was expected to have much effect on their behavior.

The analysis approach for comparing these teacher ratings for the TN-VPK participants and nonparticipants was analogous to that described above for analysis of the achievement outcomes. Multilevel models were used with children nested in their school-

level randomized applicant lists with those lists nested in districts. The same covariates were used with two exceptions. The Woodcock Johnson Composite6 baseline achievement measure was used in place of a pretest (the TN-VPK nonparticipants were not in school at baseline, thus no teacher pretest ratings were possible). In addition, a variable representing the timing of the ratings was included as a covariate, specifically the number of days between September 1 of the pre-k year and the date on which the kindergarten teacher completed the ratings. Table 7 shows the full model for the analysis of the teachers' ratings of how well prepared the children were for kindergarten participation.

Table 7: Full Analysis Results for the Kindergarten Teachers' Ratings of How Well Prepared the Children were For Kindergarten

	Coefficient	Standard error	t-value	p-value
Intercept	-16.5	1.14	-14.57	.000
Propensity score	.404	.203	1.99	.046
Rating time lag	-.001	.001	-.60	.547
Composite6 pretest	.052	.003	19.28	.000
Age (years)	.043	.138	.31	.754
Gender (1=male)	-.171	.075	-2.28	.023
Race/ethnicity Black	.166	.101	1.65	.100
Hispanic, native English	.336	.221	1.52	.129
Hispanic, not native English	.903	.133	6.78	.000
Not Hispanic, not native English	.479	.198	2.42	.016
Home literacy index	-.013	.020	-.67	.506
Mother's education	.021	.056	.38	.703
Number of working parents	-.035	.062	-.57	.569
TN-VPK participation	**.305**	**.109**	**2.79**	**.005**

Notes: Age on Sept. 1 of pre-k year; Home literacy index = sum of the z-scores for Library card, Newspaper subscriptions, and Magazine subscriptions; Mother's education (1=less than high school, 2=high school diploma/GED, 3=associate's degree, 4=more than associate's degree).

As the results in Table 7 show, there was a statistically significant difference between the TN-VPK participants and nonparticipants on the kindergarten teachers' ratings of preparedness for kindergarten, with the TN-VPK participants rated as more prepared. Table 8 provides a summary of the results from analyses parallel to this one for all the ratings made by the kindergarten teachers, including the standardized mean difference effect sizes for the contrast on these outcomes between the TN-VPK participants and nonparticipants.

Table 8: TN-VPK Effect Estimates for Kindergarten Teachers' Ratings

Outcome	TN-VPK effect estimate	p-value	Effect size
ACBR Preparedness for K (range 1-7)	.30	.005	.22
ACBR Peer Relations (range 1-7)	.04	.684	.04
ACBR Behavior Problems[a] (range 0-1)	-.01	.757	-.04
ACBR Feelings About School[a] (0-1)	-.00	.767	-.03
Cooper-Farran Social Behavior (range 1-7)	.17	.049	.19
Cooper-Farran Work-Related Skills (range 1-7)	.22	.016	.20

(a) Ratings on these scales were skewed; the analysis was done on log transformed values and those are the results shown here

The results summarized in Table 8 indicate that children who participated in TN-VPK were rated upon kindergarten entry as not only being more ready for school but also having better social behavior and better work-related skills in the classroom. Teachers did not see significant differences between the two groups in terms of their peer relations, behavior problems, or feelings about school. The implication of these findings is that the effects of exposure to TN-VPK were apparent in several ways to kindergarten teachers, and in the areas that are more closely aligned with typical focus of pre-k programs. Additionally, because effects were seen for some outcomes and not others, we have some confidence that teachers were discriminating in their ratings as opposed to possibly knowing which children were in pre-k and rating TN-VPK attenders higher across the board because of positive opinions about the program.

TN-VPK Effects for Different Subgroups of Children at the End of the Pre-K Year

As reported above, there were positive and statistically significant overall effects of TN-VPK on all but one of the WJ achievement measures examined and several of the rating scales completed by the kindergarten teachers early in the school year. These findings motivate attention to our second research question, whether there are differential effects for different subgroups of children and, if so, what subgroups show larger or smaller effects. This question was addressed by investigating the extent to which membership in various subgroups of children moderated the TN-VPK effects observed at the end of the pre-k year. In particular, we examined entering pre-k skills, age, gender, ethnicity and native English speaker status, and three family background variables as moderators of the TN-VPK effect. This was done using the analytic models similar to those described above for assessing the main effect of TN-VPK on achievement and teacher rating outcomes respectively with the addition of interaction terms for the cross products between TN-VPK participation status and variables representing the various subgroups of children.

These analyses were done for the WJ Composite6 overall achievement variable as the outcome potentially affected. The specific variables used as moderators in these analyses were the following:

- The WJ Composite6 baseline measure, included to examine differential effects for children whose achievement performance was lower versus higher at the beginning of the pre-k year.
- Age, indexed as age on September 1 of the pre-k year for the respective cohorts.
- Gender, represented by a dummy code distinguishing boys from girls.
- Race/ethnicity and whether children were native English speakers or not. The race/ethnicity of the children and whether they were native English speakers were not entirely distinct categories because most of the non-native English speaking children were Hispanic. A more differentiated set of subgroup dummy codes was therefore defined for these analyses as follows:
    - Black native English speakers (N=233)
    - Hispanic native English speakers (N=34)
    - Children for whom English is a second language irrespective of race/ethnicity (N=215)

  The remaining 594 children were White with a sprinkling of Asian and others and all native English speakers. This category was used as the reference value for the moderator variables above.
- Family background, including the home literacy index, mother's education, and number of working parents.

The initial results of analyses estimating effects on the WJ Composite6 overall achievement composite with each of these moderators included in turn showed statistically significant interactions with baseline achievement level, the home literacy index, mother's education, and English as a second language (ESL) children. Further exploration of combinations of these moderators, however, revealed that these results were being driven by interactions involving mothers' education and ESL children, particularly mothers with less than a high school education.

To more clearly reveal the nature of these interactions, the effects of TN-VPK were examined in relation to whether children were ESL or not and whether their mothers had less than a high school education versus high school or higher. These breakouts with the differences between the TN-VPK participants and nonparticipants on the WJ Composite6 achievement measure for each group along with the corresponding effect sizes are shown in Table 9. For comparability across groups and with the overall effects on the Composite6 measure reported in Table 6 earlier, these effect sizes are all standardized on the pooled standard deviations for the overall participant and nonparticipant groups.

What the summary in Table 9 reveals is that TN-VPK effects on overall achievement were much larger for ESL children than for native English speaking children (effect sizes of

.67 vs. .23). Additionally, TN-VPK effects were larger for children of mothers with less than a high school education than for children of more educated mothers (effect sizes of .53 vs. .27). Moreover, the effect size was even larger for the ESL children whose mothers had less than a high school education (ES= .88). The largest subgroup, native English speaking children with mothers who had a high school or higher education, included 74% of the total sample and had the smallest effect size of all (ES= .22).

Table 9: TN-VPK Effects on the WJ Composite6 Achievement Composite for Subgroups of Children Who Differ by English Speaking Status and Mothers' Education

	Mother's education	
Child	Less than HS (N=178) T-C diff= 8.74* Effect size= .53	HS or more (N=898) T-C diff= 4.50* Effect size= .27
English as second language (N=215) T-C difference= 11.07* Effect size= .67	T-C diff= 14.57* Effect size= .88 (N=76)	T-C diff= 9.04* Effect size= .55 (N=139)
Native English speaker (N=861) T-C difference= 3.74* Effect size= .23	T-C diff= 4.48 Effect size= .27 (N=102)	T-C diff= 3.63* Effect size= .22 (N=759)

T= TN-VPK participants; C=nonparticipants.
* p <.05

Teacher ratings. The same moderator variables identified above in relation to the WJ Composite6 outcomes were also analyzed with the ratings by the kindergarten teachers as the outcome variables. Thus each moderator variable was included in the multilevel models used to analyze the main effects on teacher ratings reported earlier in the form of an interaction term for the cross product between the centered moderator variable and the TN-VPK participant condition. The teacher ratings used as outcome variables in these analyses included all those shown in Table 8 above.

These analyses found only a few statistically significant moderator relationships. Differential TN-VPK effects were found on the ACBR Peer Relations scale for children whose mothers had less than a high school education compared with children of mothers who had completed high school or beyond. The kindergarten teachers gave somewhat higher ratings to the children with the less educated mothers (ES= .12). The number of working parents of the children being rated (a variable that took three values: 0, 1, or 2) showed significant interactions with the TN-VPK participation variable for teachers' ratings on the ACBR Preparedness for K scale, the ACBR Peer Relations scale, the Cooper-Farran Social Behavior scale, and the Cooper-Farran Work-Related Skills scale. This variable is potentially confounded with children's pre-k participation itself, which may make

employment more possible for mothers and thus is difficult to interpret in any way that has implications for identifying important differential effects on teachers' perceptions.

Whether TN-VPK Effects were Sustained through Later School Years

The results described above demonstrate positive TN-VPK effects at the end of the pre-k year on nearly all of the outcome variables included in this study. Given such favorable pre-k results, the next question is the extent to which they are sustained beyond the pre-k year. The children in this intensive substudy sample were assessed on the same WJ achievement scales annually through the end of third grade, with two more scales added at the end of the kindergarten year—Passage Comprehension and Math Calculation. In addition, first, second, and third grade teachers rated each child in the sample on the ACBR and Cooper-Farran scales at the end of each grade year.

Analysis of TN-VPK effects on these follow-up measures used the same multilevel models, propensity scores, and covariates employed in the analysis of the end of pre-k effects described above with only minor variations (e.g., dropping the rating time lag covariate that applied only to teacher ratings at the beginning of kindergarten). The WJ Passage Comprehension and Math Calculation measures added at the end of kindergarten did not have baseline pretest measures to use as a covariate as was included for the other WJ achievement measures. The baseline WJ Composite6 measure was therefore used in place of those absent pretests.

Table 10 shows the results of the analysis of the effects on the WJ achievement scales at the end of the kindergarten, and the first, second, and third grade years, with the end of pre-k results repeated for ease of comparison. In contrast to the effects found at the end of pre-k, there were no statistically significant differences between the TN-VPK participants and nonparticipants on any of these achievement measures at the end of kindergarten or at the end of first grade.

Even more striking are the effects shown on these measures at the end of the second and third grade years. During those years the benefits previously seen for the children who attended TN-VPK was reversed for all the scales, reaching statistical significance for the WJ Composite6 and Composite8 summary measures as well as several of the individual scales, most notably those assessing math achievement. That is, the children who had not attended TN-VPK outperformed the children who had attended on these measures.

The nature and magnitude of this pattern of early positive TN-VPK effects during the pre-k year that rapidly fade, then reverse, can be seen in Figure 1 where the WJ Composite6 W score outcomes are plotted for each year for each group. As Figure 1 shows, both the TN-VPK participants and nonparticipants made achievement gains each year in upward trajectories. The early advantage of the TN-VPK children disappears, however, as the nonparticipating children catch up during the kindergarten year and match the

performance of the TN-VPK participants through the end of first grade, then edge ahead in second and third grade.

Table 10: TN-VPK Effect Estimates for the Kindergarten through Third Grade Years on the Woodcock Johnson Achievement Measures

	End of pre-k year		End of kindergarten year		End of first grade year		End of second grade year		End of third grade year	
Outcome	Effect estimate	Effect size	Effect estimate	Effect size	Effect estimate	Effect size	Effect estimate	Effect size	Effect estimate	Effect size
WJ Composite6	5.32**	.32	.25	.02	-.51	-.04	-2.07*	-.15	-1.83†	-.13
WJ Composite8	N/A	-	-.13	-.01	-.70	-.05	-1.91*	-.15	-1.73†	-.13
Literacy										
Letter-Word ID	10.77**	.41	-.27	-.01	-1.56	-.05	-3.24	-.13	-3.46	-.14
Spelling	7.22**	.29	-.68	-.03	-2.11	-.10	-2.45	-.12	-2.36	-.12
Language										
Oral Comprehension	1.50†	.09	.94	.06	-.90	-.07	-1.43	-.11	-.51	-.04
Picture Vocabulary	3.66**	.20	1.01	.09	.95	.08	-.48	-.04	.77	.07
Passage Comprehension	N/A	-	-2.26	-.10	-1.61	-.08	-2.10†	-.13	-1.13	-.07
Math										
Applied Problems	4.03**	.17	1.17	.07	.55	.04	-2.38†	-.14	-3.76*	-.21
Quantitative Concepts	4.32**	.27	-1.07	-.08	-1.33	-.10	-3.45**	-.25	-2.02†	-.15
Calculation	N/A	-	-.13	-.01	-.70	-.05	-1.91*	-.15	-1.73†	-.13

Notes: Effect estimates are the coefficients on the TN-VPK participation variable indicating the difference between the mean outcomes for T-VPK participants and nonparticipants in W-score units. Effect sizes are those coefficients divided by the pooled participant and nonparticipant group standard deviations on the outcome variable.
**p<.01, *p<.05, †p<.10

A different frame of reference is provided for the achievement trajectories of the TN-VPK participants and nonparticipants when the WJ standard scores are examined in place of the longitudinally scaled W-scores. The standard scores are normed so that a score of 100 represents the mean score for the norming sample, presumed to be representative of the national population of children at each respective age. Figure 2 shows these standardized scores from the pre-k year through third grade for the TN-VPK participants and nonparticipants.

The pattern of achievement gains when scores are referenced to the test norms is rather different from that seen in Figure 1. As in Figure 1, the TN-VPK participants show greater gains during the pre-k year than nonparticipants, with nonparticipants catching up

Figure 1: W-Scores on WJ Composite6 for the TN-VPK Participant and Non-Participant Groups on Each Wave of Measurement

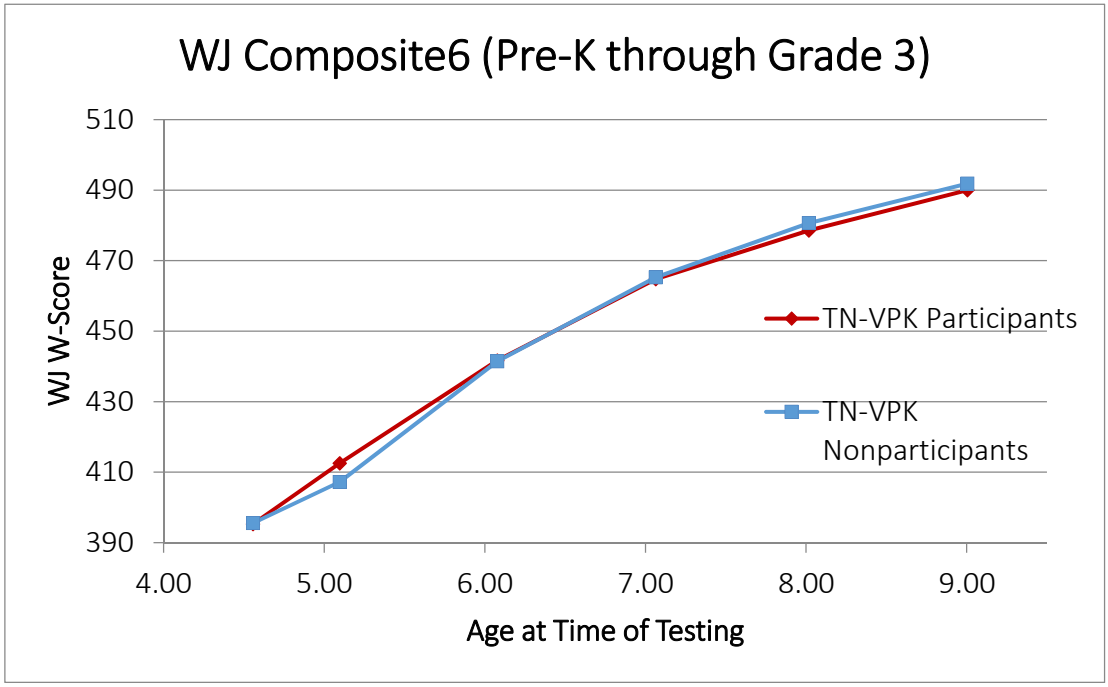
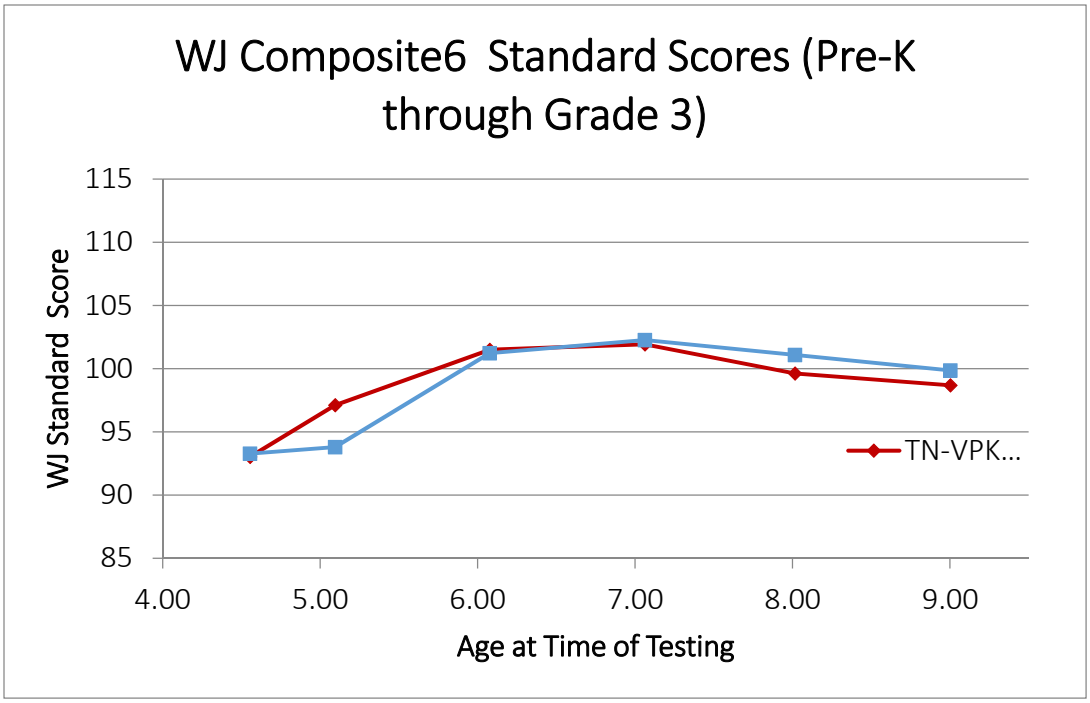


Figure 2: Standard Scores on WJ Composite6 for the TN-VPK Participant and Non-Participant Groups on Each Wave of Measurement



in kindergarten and outperforming the participants in second and third grade. In addition, however, Figure 2 shows that, relative to national norms, the early gains made by both groups begin to flatten out in first grade and actually turn downward in second and third grade.

Moderator relationships with follow-up achievement Outcomes. The analysis of TN-VPK effects at the end of pre-k reported earlier identified two significant moderators of those effects as indexed by the WJ Composite6 measure. Larger effects were found for children for whom English was a second language than for children who were native English speakers. Further, larger effects were found for children of mothers' with less than a high school education than for children who completed high school or more. The analysis of the follow-up waves of outcome measures thus also included an examination of the three-way interaction between these moderators and TN-VPK participation shown earlier in Table 9, but no significant effects were found. The difference in baseline scores on the WJ Composite6 measure was especially large for the children with English as a second language compared with native English speaking children, however.

In light of the overall finding of no difference between TN-VPK participants and nonparticipants on the WJ achievement measures by the end of kindergarten with effects reversing in second and third grade, it is informative to consider whether that same pattern characterizes the native English speaking and ESL children, recognizing that there are increasing proportions of ESL children in Tennessee classrooms. Table 11 reports the mean W-scores on the Composite6 outcomes from baseline to end of third grade for these two groups of children, further divided into TN-VPK participants and nonparticipants. The mean observed scores are reported for the TN-VPK participant groups and means that are covariate adjusted to match the characteristics of the respective participants are reported for the nonparticipant groups. The only statistically significant interaction between native language status and TN-VPK participation was the one that occurred at the end of the pre-k year and was described earlier, but the large baseline differences are evident.

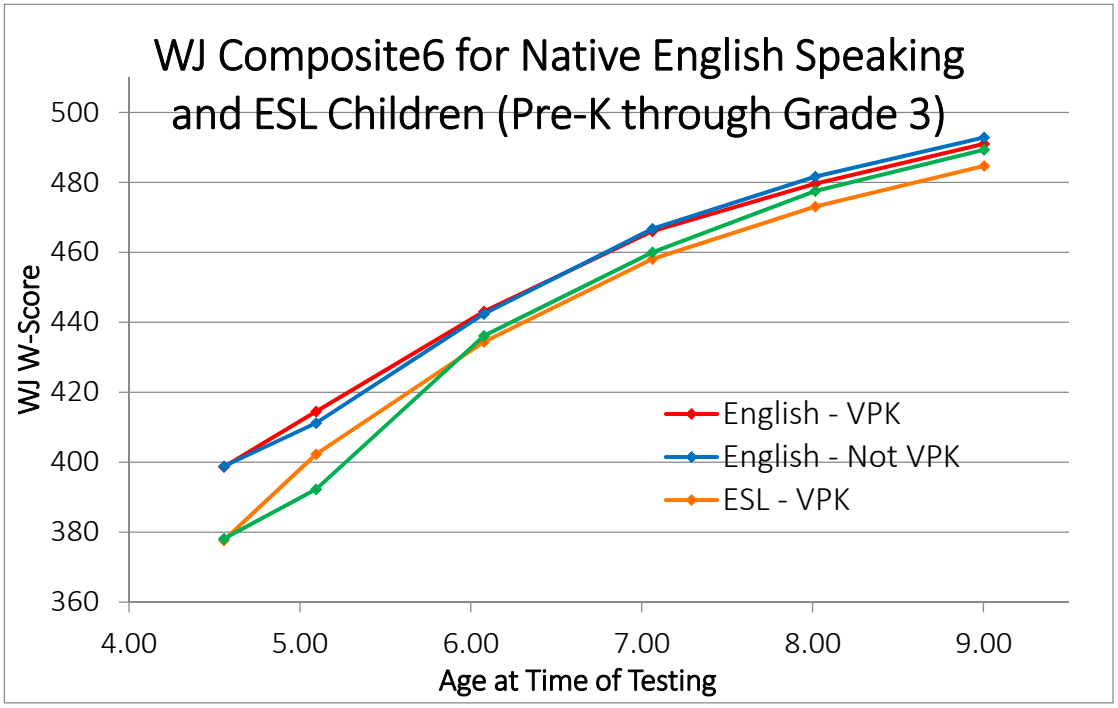
Table 11: ESL-Native English Moderator of Effects on WJ Composite6

Language	TN-VPK	Baseline	End of pre-k*	End of k	End of 1st grade	End of 2nd grade	End of 3rd grade
Native English	Yes	398.7	414.5	443.1	466.1	479.6	491.1
	No	398.8	411.2	442.4	466.7	481.6	492.9
English as Second Language	Yes	377.7	402.3	434.4	458.1	473.1	484.7
	No	378.1	392.2	436.1	460.0	477.5	489.3

* $p < .05$ for the Language x TN-VPK participation condition interaction term in the regression model.

Figure 3 shows the trajectories on the Composite6 W-scores for the ESL vs native English speakers from baseline through the end of third grade. The lower starting point and especially strong gains made by the ESL children during the pre-k year can be clearly seen there. As with the overall sample, however, the TN-VPK advantage has disappeared for them by the end of the kindergarten year and the TN-VPK nonparticipants begin outperforming the participants after that. Perhaps most striking in Figure 3, however, is the performance of the ESL children in the later grades. Though they began with much lower achievement scores than the native English speaking children, they had closed much of that gap by the end of kindergarten and, for the TN-VPK nonparticipants, even more of it by the end of third grade. The native English speaking children, by contrast, showed much smaller effects of TN-VPK participation and much smaller differences between the TN-VPK participants and nonparticipants at the end of the second and third grade years.

Figure 3: WJ Composite6 for ESL and Native English Speakers at Each Wave of Measurement Broken out by TN-VPK Participation



Teacher ratings. As with the achievement measures at the end of the pre-k year, the ratings by kindergarten teachers at the beginning of the kindergarten year showed several positive effects of TN-VPK participation and no adverse effects. The results of the analysis of the teacher ratings at the end of first, second, and third grade on the same rating scales are shown in Table 12, along with those reported earlier for the beginning of kindergarten for ease of comparison. Here again, as with the achievement measures, some of the positive effects found after the pre-k year have reversed. At the end of the first grade

year, teachers rate the TN-VPK participants significantly lower than participants on work-related skills, feelings about school, and preparedness for grade. Indeed, all of the effect estimates have turned negative, though only those three reach statistically significant levels (marginally for preparedness for grade). However, by the end of the second grade there are no longer any significant differences between TN-VPK participants and nonparticipants, and that pattern continues into third grade with the exception of a marginally significant positive effect for the TN-VPK participants on the teachers' ratings of peer relations.

Table 12: TN-VPK Effect Estimates for First, Second, and Third Grade Teachers' Ratings

	Start of kindergarten year		End of first grade year		End of second grade year		End of third grade year	
Outcome	Effect estimate	Effect size	Effect estimate	Effect size	Effect estimate	Effect size	Effect estimate	Effect size
ACBR Preparedness for Grade	.30*	.22	-.24†	-.17	.07	.05	-.01	-.01
ACBR Peer Relations	.04	.04	-.05	-.05	.04	.04	.21†	.19
ACBR Behavior Problems	-.008	-.04	-.004	-.02	-.016	-.07	-.039	-.16
ACBR Feelings About School	-.002	-.03	-.014*	-.21	.003	.04	.002	.03
CF Social Behavior	.17*	.19	-.15	-.16	.06	.06	.07	.07
CF Work-Related Skills	.22*	.20	-.24*	-.20	.00	-.00	.10	.08

Notes. Scoring range on scales: ACBR Preparedness (1-7); ACBR Peer Relations (1-7); ACBR Behavior Problems (log transformed, 0-1); ACBR Feelings About School (log transformed, 0-1); Cooper-Farran Social Behavior (1-7); Cooper-Farran Work-Related Skills (1-7).
*p<.05, †p<.10

The only moderator of teachers' ratings found at the end of the pre-k year (described earlier) was the baseline variable from the parent survey that asked about parental employment. As noted earlier, this variable is potentially confounded with the pre-k status of the children and the implications of these statistical interactions are thus unclear.

Discussion

Summary of Findings

Results from this randomized control trial of a state funded targeted pre-k program delivered at scale are complex. We were able to assess a subset of a large randomized sample for 1076 children whose parents provided consent for annual data collection from those children. This intensive substudy sample included 773 children who participated in Tennessee's Voluntary Pre-K program and 303 children who did not attend because there was not space for them in the oversubscribed programs participating in the randomization. The characteristics of the children in these two groups were quite similar at baseline and, to further ensure that they were comparable, selected baseline covariates, including propensity scores, were used as statistical controls in all analyses. The TN-VPK participants attended pre-k classes for an average of 147 days during the pre-k year. Most of the children in the control group were cared for at home, although about 27% attended Head Start or a private childcare center.

Children were individually assessed on a variety of achievement tests measuring aspects of school readiness, including literacy, language and math. These tests were administered at the beginning and end of the pre-k year and annually thereafter at the end of each grade year through third grade. In addition to the achievement measures, children's behaviors of a sort that some call "non-cognitive" were rated annually by their teachers. The first ratings were obtained at the beginning of kindergarten when all children had entered school; those ratings are considered an evaluation of experiences during the pre-k year. Thereafter, first through third grade teachers rated children's behaviors each spring.

Effects at the end of pre-k. Our research focused on three primary questions. The first concerned the effectiveness of the TN-VPK program for preparing children for kindergarten entry. We found that, at the end of pre-k, the TN-VPK children had significantly higher achievement scores on all six of the achievement subtests administered, with the largest effects on the two literacy measures. The effect size on the composite achievement measure that combined the scores on all six measures was .32. This effect is of the same magnitude Duncan and Magnuson (2013) reported for end of year effects for the pre-k programs included in their comprehensive research review and is larger than the average for programs enacted since the 1980s. Also, at the beginning of kindergarten, the teachers rated the TN-VPK children as better prepared for kindergarten work, as having better behaviors related to learning in the classroom, and as having more positive peer relations. They did not see the children as having more behavior problems and both groups of children were rated as being highly positive about school.

Differential pre-k effects. The second question our research addressed was whether some identifiable subgroups of children were differentially affected by TN-VPK attendance. We examined a number of relevant moderators of the pre-k effects and found

no differences for gender, ethnicity, or age of enrollment. The moderators we did find were driven by the relationship of mothers' education and children for whom English was a second language to the magnitude of the TN-VPK effects. The TN-VPK effects were the largest for children who were learning English and whose mothers had less than a high school degree. English language learners with more educated mothers had the next largest effect size. The effects for native English speakers whether or not their mothers had a high school degree were considerably smaller.

Persistence of pre-k effects. The third question we addressed involved the sustainability of effects on achievement and behavior beyond kindergarten entry and through the third grade year. The children who participated in TN-VPK and the control group of children who did not participate were followed and reassessed in the spring every year, with more than 90% of the initial sample located and included each year. By the end of kindergarten, the control children had caught up to the TN-VPK children and there were no longer significant differences between them on any achievement measure. Thus the control children gained as much in one year on these achievement tests as the TN-VPK children had in two years. The same result was obtained at the end of first grade—no differences between the TN-VPK participants and nonparticipants on the achievement measures.

By the end of the second grade year, however, the groups began to diverge with the TN-VPK children scoring somewhat lower than the control children on most of the achievement measures. These differences were statistically significant for both the achievement composite measures and the math subtests. The moderating effects of ESL status and mothers' education were no longer significant, but it is interesting to note that, whether or not the ESL children participated in TN-VPK, by the end of third grade their achievement scores were higher than those of either the native English speaking TN-VPK participants or nonparticipants.

In terms of behavioral effects, by the spring of the first grade year, teachers rated the TN-VPK children as less well prepared for school, having poorer work skills, and feeling more negative about school. This was a reversal of the ratings provided by the kindergarten teachers at the beginning of kindergarten. It is notable that these ratings precede the downward achievement trend for TN-VPK children that appeared in the second and third grades.

Implications

Our findings on the follow-up effects of TN-VPK participation were unexpected. We interpret them cautiously recognizing, as distinguished evaluation researchers have noted, that no single study, no matter how carefully done, produces definitive results (Campbell, 1969; Cook 2003). But we would also note that, just because the results of an evaluation do not support a currently popular view, it does not mean that they are wrong. In a review of

social policy studies in the U.K., Ettelt, Mays, and Allen (2015) observed that when evaluation findings turned out not to support current policy, they tended to be ignored "or, worse, purposely misinterpreted" (p. 294).

Much of the expectation for long-term positive pre-k effects comes from the small experimental studies of model programs conducted 40 to 50 years ago that were discussed at the beginning of this report. The results of those studies continue to be cited as the reason businesses and the government should invest in pre-kindergarten programs (e.g., Christeson, Bishop-Josef, O'Dell-Archer, Beakey; & Clifford, 2013; Kay, & Pennucci, 2014; ReadyNation, nd; President's Council of Economic Advisors, 2014). But we are also led to expect benefits from pre-k intervention by more recent research that frequently finds positive effects of public pre-k programs at the end of the pre-k year with an associated implied expectation that they would be sustained to some degree.

The studies that have investigated longer term effects have generally used weaker matched designs rather than randomized designs, but their results are not so different from those we have reported here—typically a "fade out" of the initial effects with, perhaps, small but usually nonsignificant differences favoring the pre-k group on some measures. Exceptions are two matching studies (Deming, 2009; Reynolds et al., 2011) that found achievement effects past second grade. As we indicated earlier, however, the difficulty of matching groups on the interest of parents in enrolling their child in a pre-k program makes the interpretation of all the matching studies uncertain.

A more appropriate comparison is with the recent Head Start Impact study, which like this TN-VPK study, is a prospective, random assignment study of a publicly funded pre-k program (Puma et al., 2012). The Head Start Impact study was broader than the TN-VPK study, focusing as it did on a nationally implemented program. Nonetheless, the results were similar. The Impact study found positive effects at the end of the pre-k year, with the largest effects on the literacy measures and smaller effects on math, just as in this TN-VPK study. It also found that those effects did not persist past the end of kindergarten with only limited exceptions. Puma et al. were as perplexed by their findings as we are by ours:

> Although the underlying cause of the rapid attenuation of early impacts is an area of frequent speculation, we don't have a good understanding of this observed pattern. All we can say is after the initially realized cognitive benefits for the Head Start children, these gains were quickly made up by children in the non-Head Start group (p. 151).

These findings have led us to think about many dimensions of implementing scaled up publicly funded pre-k programs, some of which will be discussed in the next sections.

Defining "pre-k". The TN-VPK program is similar to other new pre-k initiatives in that its classrooms are primarily located in public schools, in effect adding a grade below kindergarten. This way of organizing pre-k programs is one that is supported, for instance, by the new federal pre-k expansion initiative (Federal Register, 2014); the funding for

expanding or developing pre-k programs is funneled through the local education agencies (LEA), as is TN-VPK. However, this is not the only way states have gone about providing early intervention experiences for children.

Some states like Florida rely entirely on private providers, giving families a voucher they can use at approved programs. A recent debate in Minnesota was won by proponents of scholarships for low-income families to purchase care in the market place. In North Carolina, Smart Start, begun by Governor Hunt in the early 1990s, did not focus on classrooms at all. Instead, funding was allocated to counties to create higher quality and seamless services for children aged 0-5 within the county, and it was left up to the counties to determine how to do that (Ladd, Muschkin & Dodge, 2014). The Division of Child Development and Early Education in the NC Department of Health and Human Services, not the Department of Education, oversees Smart Start and its offshoot, More at Four.

As Quinton (2014) recently noted: "…while there's a growing consensus on the value of preschool, states disagree on where the programs should be based, who should run them, or how the government should support them" (p.2). This circumstance makes generalizations about the results from evaluations of statewide programs problematic. Our study is most relevant to programs housed in elementary schools and overseen by the state departments of education. Other types of preschool and early childhood programs may produce different effects.

***Determining quality*.** Another issue is program quality. When Tennessee began its voluntary pre-k program, it looked for guidance, as many states do, to the benchmarks established by the National Institute for Early Education Research (Barnett, et al, 2014). TN-VPK meets 9 of those 10 benchmarks and is among the states meeting most of those benchmarks. In the recent request for applications for preschool expansion grants from the U.S. Department of Education, the term "high-quality" pre-k is used throughout, and defined mainly in terms of these same NIEER benchmarks. Our findings for the TN-VPK program raises questions about whether those benchmarks prescribe elements of pre-k programs that are linked to long term positive effects on either achievement or behavior (Mashburn et al., 2010).

Over the past 30 or so years, there have been many attempts to define what high quality means for preschool and now pre-kindergarten classrooms (see Farran & Hofer, 2013, for a review). Many states rely on rating systems to determine the quality of their early childhood classrooms, e.g. the *Early Childhood Environmental Rating Scale* (ECERS; Harms & Clifford, 1980; with several further editions) or the *Classroom Assessment Scoring System* (CLASS; LaParo & Pianta, 2003) now required of Head Start classrooms. Recently, Weiland, Ulvestad, Sachs, and Yoshikawa (2013) examined ratings from the CLASS and ECERS in relation to the outcomes in the Boston public pre-k program. They concluded that classroom quality as measured by these instruments had no or very small relationships to children's gains in developmental outcomes, even when they used the threshold analysis

suggested by Burchinal, Vanderbrift, Pianta, and Mashburn (2010). Weiland et al. argue that these measures were simply not strong indices of quality.

If we are to continue offering pre-k through the public school system, fundamental empirical work may be required to identify specific behaviors and instructional practices important for young children's development in that environment. For example a recent study involving 60 pre-k classrooms in elementary schools demonstrated that the emotional tone, quality of instruction, and level of child involvement in math and literacy activities were significant factors in predicting gains in self-regulation over the year (Fuhs, Farran, & Nesbitt, 2013). States need guidance beyond what is presently available in order to establish pre-k classrooms that indeed have "high-quality" and positive outcomes.

***Alignment with K-3*.** Our findings highlight the importance of the K-3rd grade experience for children, especially children from low-income backgrounds. The fade out of pre-k effects could, at least in part, be due to failure of kindergarten teachers to build on the skills children bring with them from their pre-k experiences. This might happen, for example, if they are mainly directing their attention to the children who need it the most, thus allowing them to catch up with those who have been in pre-k. This is an empirical question that we do not have the data to address. Nonetheless some explorations of what kindergarten teachers are covering in their classrooms suggests that they may be out of touch with the skills their children possess (Claessens, Engel, & Curran, 2014). Claessens et al. found that higher levels of instruction in math and literacy benefited all children in the class, regardless of preschool experience. Thus it may not be that the teachers are teaching specifically to the children with the greatest need; rather, it may be that their instruction has not caught up to what all the children are prepared to learn.

Children from TN-VPK classrooms and their counterparts in the control group were eligible for TN-VPK because their families were impoverished. After pre-k, these children tended to attend high poverty schools. Of concern from our findings is the fact that the achievement of both the TN-VPK participants and nonparticipants begins to decline in second and third grades. Reardon (2011) has rightfully called attention to the widening achievement gap between the rich and the poor, and thus it is important to determine when that actually begins. Our data suggest that these children from economically disadvantaged families were very responsive to their introduction to formal schooling in kindergarten whether or not they had participated in TN-VPK beforehand. But their momentum was not maintained by their instructional experiences in first through third grade; in fact, quite the reverse. Halpern (2013) rightfully cautions against making early childhood education less "early-childhood-like" (p. 23), speaking to the pressure to make pre-k classrooms more and more academic; we might also need to focus on making the full K-3 instructional spectrum richer and more instructionally deep.

Conclusion

As we noted at the beginning of this paper, increasing numbers of children are living in impoverished circumstances, circumstances that have immediate and long lasting consequences for them. Pre-k intervention has been proposed as one way to address this problem and is expanding quickly in many states and with federal endorsement. However, the idea that pre-k can be scaled up quickly, cheaply, and without professional support or vision is certainly bound to be incorrect. Assumptions about what poor children are experiencing in their families lead to comments like: "… even a lower-quality preschool program can have an impact on children from the most disadvantaged environments" (Cascio & Schanzenbach, 2014, p. 2). But it is not at all obvious that the rush to implement pre-k programs widely without the necessary attention to the quality of the program provides worthwhile benefits to children living in those disadvantaged environments. As Kirp (2009) cautioned, scaling up pre-k programs quickly could lead to badly run programs that might, in fact, be worse than doing nothing.

The TN-VPK program saturates the state; every county has at least one classroom and all school districts except one have endorsed the program by opening new classrooms. Thus, the structural support exists in the state to continue to explore pre-k as a means for preparing children for success in school, but we need to think carefully about what the next steps should be. It is apparent that the term pre-k, or even "high-quality" pre-k, does not convey actionable information about what the critical elements of the program should be. Now is the time to pay careful attention to the challenge of serving the country's youngest and most vulnerable children well in the pre-k programs that have been developed and promoted with their needs in mind.

References

Almond, D. & Currie, J. (2010). *Human capital development before age five.* National Bureau of Economic Research, Working Paper 15827, http://www.nber.org/papers/w15827

Andrews, R., Jargowsky, P. & Kuhne, K. (December, 2012). *The effects of Texas's targeted pre-kindergarten program on academic performance.* Working Paper, 18598. National Bureau of Economic Research, Cambridge, MA. http://www.nber.org/papers/w18598

Baker, M. (2011). Innis Lecture: Universal early childhood interventions: What is the evidence base? *Canadian Journal of Economics/Revue canadienne d'économique*, *44*(4), 1069-1105.

Bania, N., Kay, N., Aos, S., & Pennucci, A. (2014). Outcome evaluation of Washington State's *Early Childhood Education and Assistance Program,* (Document No. 14-12-2201). Olympia: Washington State Institute for Public Policy.

Barnett, S., Carolan, M., Squires, J., Brown, K., & Horowitz, M. (2014). *The state of preschool 2014: State preschool yearbook.* National Institute for Early Education Research (NIEER), Graduate School of Education, Rutgers, the State University of New Jersey.

Bartik, T., Gormley, W., & Adelstein, S. (2011). Earnings benefits of Tulsa's program for different income groups. Upjohn Institute Working Paper 11-176. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

Burchinal, M., Vanderbrift, N., Pianta, R., & Mashburn, A. (2010). Threshold analysis of associations between child care quality and child outcomes for low-income children in pre-kindergarten programs. *Early Childhood Research Quarterly, 25,* 166-176. doi:10.1016/j.ecresq.2009.10.004

Campbell, D. (1969). Reforms as experiments. *American Psychologist, 24,* 409-429.

Campbell, F. A., Pungello, E. P., Miller-Johnson, S., Burchinal, M., & Ramey, C. T. (2001). The development of cognitive and academic abilities: Growth curves from an early childhood educational experiment. *Developmental Psychology, 37(2),* 231-242. doi: 10.1037/0012-1649.37.2.231

Cascio, E. U. & Schanzenbach, D. W. (2013). *The impacts of expanding access to high quality preschool education*. Paper presented at the fall 2013 Brookings Panel on Economic Activity.

Cascio, E.U. & Schanzenbach, D. (2014). Proposal 1: Expanding preschool access for disadvantaged children. In M. Kearney & B. Harris (Eds.), *The Hamilton Project: Policies to address poverty in America.* Washington, DC: Brookings.

Claessens, A., Engel, M., Curran, F.C. (2014). Academic content, student learning and the persistence of preschool effects. *American Educational Research Journal, 51,* 403-434. DOI: 10.3102/0002831213513634

Cook, T. (2003). Why have educational evaluators chosen not to do randomized experiments? The ANNALS of the American Academy of Political and Social Science, 589, 114-149.

Currie, J. & Rossin-Slater, M. (2014). Early-life origins of life-cycle well-being: Research and policy implications. *Journal of Policy Analysis and Management, 34,* 208-242. DOI:10.1002/pam.21805.

Deming, D. (2009). Early childhood intervention and life-cycle skill development: Evidence from Head Start. *American Economic Journal: Applied Economics, 1:3,* 111-134.

Duncan, G. & Magnuson, K. (2013). Investing in preschool programs. *Journal of Economic Perspectives, 27,* 109-132. Dx.doi.org/10.1257/jeb27.2.109.
Duncan, G., Magnuson, K., & Votruba-Drzal, E. (2014). Boosting family income to promote child development. *The Future of Children, 24,* 99-120.
Duncan, G., Ziol-Guest, K., & Kalil, A. (2010). Early childhood poverty and adult attainment, behavior, and health. *Child Development*, 81, 306–325, doi: 10.1111/j.1467-8624.2009.01396.x.
Ettelt, S., Mays, N., & Allen, P. (2015). Policy experiments: Investigating effectiveness or confirming direction. *Evaluation, 21*, 292-307. DOI: 10.1177/1356389015590737.
Farran, D. C. (2007). *Is education the way out of poverty? A Reflection on the 40th anniversary of Head Start* (with commentaries by James King and Bernard L. Charles), Center for Research on Child Development and Learning, No. 3 (50 pages – ISBN: 0-9727709-2-5).
Farran, D.C. & Hofer, K. (2013). Evaluating the quality of early childhood education programs. In O. Saracho & B. Spodek (Eds.), *Handbook of Research on the Education of Young Children* (pp. 426-437). New York, NY: Routledge/Taylor & Francis.
Federal Register, 79, No. 159. (August 18, 2014*). Applications for New Awards; Preschool Development Grants – Expansion Grants.* Washington, DC*:* Department of Education and Department of Health and Human Services*.*
Fitzpatrick, M. (2008). Starting school at four: The effect of universal pre-kindergarten on children's academic achievement. *The B.E. Journal of Economic Analysis & Policy, 8,* 1-38.
Fuhs, M., Farran, D., & Nesbitt, K. (2013*)*. Preschool classroom processes as predictors of children's cognitive self-regulation skills development. *School Psychology Quarterly, 28,* 347-359. DOI: 10.1037/spq0000031
Gormley, W. T., Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal pre-K on cognitive development. *Developmental Psychology*, *41*, 872–884. doi:10.1037/0012-1649.41.6.872
Grehan, A., Cavalluzzo, L., Gnuschke, J., Hanson, R., Oliver, S, and Vosters, K. (2011). *Participation during the first four years of Tennessee's Voluntary Prekindergarten program.* (Issues & Answers Report, REL 2011-No. 107). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Appalachia. Retrieved from http://ies.ed.gov/ncee/edlabs.
Harms, T., & Clifford, R. M. (1980). *The Early Childhood Environment Rating Scale*. New York: Teachers College Press
Halpern, R. (2013). Tying early childhood education more closely to schooling: Promise, perils and practical problems. *Teachers College Record, 115,* 1-28.
Heckman, J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science, 312,* 1900-1902. http://www.jstor.org/stable/3846426
Hill, C., Gormley, W., & Adelstein, S. (2015). Do the short-term effects of a high-quality preschool program persist? *Early Childhood Research Quarterly, 32,* 60-79. dx.doi.org/10.1016/j.ecresq.2014.12.005
Huang, F., Invernizzi, M., & Drake, A. (2012). The differential effects of preschool: Evidence from Virginia. *Early Childhood Research Quarterly, 27,* 33-45. doi:10.1016/j.ecresq.2011.03.006

Huston, A., Gupta, A., & Schexnayder, D. (March, 2012). *The relationship of Pre-K attendance to 3rd grade test results.* Ray Marshall Center for the Study of Human Resources, Austin, TX.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press.

Kay, N., & Pennucci, A. (2014). *Early childhood education for low-income students: A review of the evidence and benefit-cost analysis* (Doc. No. 14-01-2201). Olympia: Washington State Institute for Public Policy.

Kirp, D. (2009). *The sandbox investment: The preschool movement and kids-first politics.* Cambridge, MA: Harvard University Press.

Ladd, H., Muschkin, C., & Dodge K. (2014). From birth to school: Early childhood initiatives and third grade outcomes in North Carolina. *Journal of Policy Analysis and Management, 33,* 162-187. DOI:10.1002/pam.21734

LaParo, K. M., & Pianta, R. C. (2003). *CLASS: Classroom Assessment Scoring System*. Charlottesville: University of Virginia

Lazar, I., Darlington, R., Murray, H., Royce, J., & Snipper, A. (1982). Lasting effects of early education: A report from the Consortium for Longitudinal Studies. *Monographs of the Society for Research in Child Development, 47*(2–3, Serial No. 195). doi:10.2307/1165938

Lipsey, M. W., Farran, D. C., Bilbrey, C., Hofer, K. G., Dong, N. (2011). *Initial Results of the Evaluation of the Tennessee Voluntary Pre-K Program* (Research Report). Nashville, TN: Vanderbilt University, Peabody Research Institute. (https://my.vanderbilt.edu/tnprekevaluation/files/2013/10/April2011_PRI_Initial_TN-VPK_ProjectResults.pdf)

Lipsey, M. W., Hofer, K. G., Dong, N., Farran, D. C., & Bilbrey, C. (2013a). *Evaluation of the Tennessee Voluntary Prekindergarten Program: End of Pre-K Results from the Randomized Control Design* (Research Report). Nashville, TN: Vanderbilt University, Peabody Research Institute. (https://my.vanderbilt.edu/tnprekevaluation/files/2013/10/May2013_PRI_EndofPK_TN-VPK_RCT_ProjectResults.pdf)

Lipsey, M. W., Hofer, K. G., Dong, N., Farran, D. C., & Bilbrey, C. (2013b). *Evaluation of the Tennessee Voluntary Prekindergarten Program: Kindergarten and First Grade Follow-Up Results from the Randomized Control Design* (Research Report). Nashville, TN: Vanderbilt University, Peabody Research Institute. (https://my.vanderbilt.edu/tnprekevaluation/files/2013/10/August2013_PRI_Kand1stFollowup_TN-VPK_RCT_ProjectResults_FullReport1.pdf)

Lipsey, M., Weiland, C., Yoshikawa, H., Wilson, S., & Hofer, K. (2015). Prekindergarten age-cutoff regression-discontinuity design: Methodological issues and implications for application. *Educational Evaluation and Policy Analysis, 37,* 296-313. DOI: 10.3102/0162373714547266

Mashburn, A., Pianta, R., Hamre, B., Downer, J., Barbarin, O., Bryant, D., … Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development, 79,* 732-749

Minervino, J. & Pianta, R. (September, 2014). Early learning: The new fact base and cost sustainability. In J. Minervino (Ed.), *Lessons from research and the classroom.* Washington: Bill & Melinda Gates Foundation.

Mistler, S. A. (2013). *A SAS® macro for applying multiple imputation to multilevel data.* Proceedings of the SAS Global Forum 2013, San Francisco, California. Contributed paper (Statistics and Data Analysis), 438-2013.

National Education Goals Panel, (archives). *History 1989 to Present.* http://govinfo.library.unt.edu/negp/page1-7.htm, retrieved September 12, 2015

Peck, L. R., &. Bell, S. (2014). *The role of program quality in determining Head Start's impact on child development.* OPRE Report #2014-10, Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Puma, M., Bell, S., Cook, R., Heid, C., Broene, P., Jenkins, D., Mashburn, A. & Downer, J. (2012). *Third grade follow-up to the Head Start Impact Study final report,* OPRE Report # 2012-45, Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Quinton, S. (September 4, 2015). *States agree on need for 'preschool,' differ on definition.* The Pew Charitable Trusts/Research & Analysis/Stateline. http://www.pewtrusts.org/en/research-and-analysis/blogs/stateline/ Retrieved September 5, 2012.

ReadyNation (n.d.). *Business case for early childhood investments.* Washington, DC: ReadyNation, www.ReadyNation.org. Retrieved September 12, 2015.

Reardon, S. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. In G. Duncan & R. Murnane, (Eds). *Whither opportunity: Rising inequality, schools, and children's life chances*. New York: Russell Sage Foundation.

Reynolds, A., Temple, J, Robertson, D., White, R., & Ou, S-R (2011). Age 26 cost-benefit analysis of the Child-Parent Center early education program. *Child Development, 82*, 379-404. DOI: 10.1111/j.1467-8624.2010.01563.x

Rosinsky, K. (2014). *The relationship between publicly funded preschool and 4th grade math test scores: A state-level analysis.* (Master's thesis, Georgetown University, Washington, DC). Retrieved from https://m.repository.library.georgetown.edu/bitstream/handle/10822/709852/Rosinsky_georgetown_0076M_12517.pdf?sequence=1&isAllowed=y, August 31, 2015

Schweinhart, L. J., Montie, J., Xiang, Z., Barnett, W. S., Belfield, C. R., & Nores, M. (2005). *Lifetime effects: the High/Scope Perry Preschool study through age 40*. Ypsilanti, MI: High/Scope Press.

Strategic Research Group. (May, 2011). *Assessing the impact of Tennessee's Pre-kindergarten program: Final report.* Columbus Ohio: Strategic Research Group.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Vogel, C.A., Xue, Y., Moiduddin, E., Kisker, E., & Carlson, B.L. (2010). *Early Head Start children in grade 5: Long-term follow-up of the Early Head Start research and evaluation project study sample.* OPRE Report #2011-8, Washington DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office for Planning, Research, and Evaluation.

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

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

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson Tests of Cognitive Abilities-III.* Itasca, IL: Riverside.

Wong, V. C., Cook, T. D., Barnett, W. S., & Jung, K. (2008). An effectiveness-based evaluation of five state pre-kindergarten programs. *Journal of Policy Analysis and Management*, *27*, 122–154. doi:10.1002/pam.20310

Zhai, FH, Brooks-Gunn, J., & Waldfogel, J. (2014). Head Start's impact is contingent on alternative type of care in comparison group. *Developmental Psychology, 50,* 2572-2586. DOI: 10.1037/a0038205