Early Childhood Research Quarterly xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

Early Childhood Research Quarterly

Effects of the Tennessee Prekindergarten Program on children's achievement and behavior through third grade

Mark W. Lipsey*, Dale C. Farran, Kelley Durkin

Vanderbilt University, United States

ARTICLE INFO

Article history:
Received 9 June 2017
Received in revised form 20 February 2018
Accepted 9 March 2018
Available online xxx

Keywords:
Public pre-k
Randomized control trial
Longitudinal
Early childhood education
Achievement
Policy

ABSTRACT

This report presents results of a randomized trial of a state prekindergarten program. Low-income children ($N = 2990$) applying to oversubscribed programs were randomly assigned to receive offers of admission or remain on a waiting list. Data from pre-k through 3rd grade were obtained from state education records; additional data were collected for a subset of children with parental consent ($N = 1076$). At the end of pre-k, pre-k participants in the consented subsample performed better than control children on a battery of achievement tests, with non-native English speakers and children scoring lowest at baseline showing the greatest gains. During the kindergarten year and thereafter, the control children caught up with the pre-k participants on those tests and generally surpassed them. Similar results appeared on the 3rd grade state achievement tests for the full randomized sample – pre-k participants did not perform as well as the control children. Teacher ratings of classroom behavior did not favor either group overall, though some negative treatment effects were seen in 1st and 2nd grade. There were differential positive pre-k effects for male and Black children on a few ratings and on attendance. Pre-k participants had lower retention rates in kindergarten that did not persist, and higher rates of school rule violations in later grades. Many pre-k participants received special education designations that remained through later years, creating higher rates than for control children. Issues raised by these findings and implications for pre-k policy are discussed.

In 2015, 67% of U.S. children 4 years old and not in kindergarten were enrolled in preschool programs (McFarland et al., 2017). As in years past, higher income families were more likely to enroll their children in some form of center-based care than low-income families, and low-income children were more likely to be enrolled in public programs such as Head Start and state-funded prekindergarten programs. Many states have been prompted to increase funding for pre-k programs in order to serve a greater number of high-risk children (Parker, Workman, & Atchison, 2016) and most states currently offer some form of voluntary pre-k that is available to children from low-income families (Barnett et al., 2017).

State funding targeted to children from low-income families implies goals beyond merely providing daycare. For example, Mississippi began state funding of pre-k in 2014 after lobbying by Mississippi First about the role pre-k can play "... in closing the achievement gap while raising achievement for all learners" (http://www.mississippifirst.org/education-policy/pre-kindergarten/). In 2014–2015, the U.S. Department of Education allocated millions of dollars to states to expand pre-k, citing a white paper asserting that high quality early education narrows achievement gaps, boosts adult earnings, and results in savings of $8.60 for every $1 spent (Executive Office of the President of the United States, 2014). With such high expectations, it is especially important for policy to be informed by research on the effects of state-funded pre-k.

1. Pre-k effects at kindergarten entry

One relevant body of evidence demonstrates that state pre-k programs generally improve such aspects of children's readiness for kindergarten as letter recognition and print awareness (Gormley, Gayer, Phillips, & Dawson, 2005; Wong, Cook, Barnett, & Jung, 2008). Most of what is known about these immediate pre-k effects comes from age-cutoff regression-discontinuity designs (RDD). Though not without potential biases (Lipsey, Weiland, Yoshikawa, Wilson, & Hofer, 2015), this design has the twofold advantage of being recognized as a relatively strong design while also being easily applied to any program with an age cutoff for admission. Chil-

* Corresponding author at: Peabody Research Institute, Vanderbilt University, 230 Appleton Place, PMB 181, Nashville, TN 37203-5721, United States.
E-mail addresses: mark.lipsey@vanderbilt.edu (M.W. Lipsey), dale.farran@vanderbilt.edu (D.C. Farran), kelley.durkin@vanderbilt.edu (K. Durkin).

https://doi.org/10.1016/j.ecresq.2018.03.005

dren 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 be measured after the admitted group completes pre-k, and the group in waiting is just beginning, and then compared with statistical adjustments for the age difference.

The age-cutoff RDD was first used in an evaluation of the Tulsa pre-k program that found positive effects (Gormley et al., 2005). A number of similar age-cutoff RDD studies have been conducted since and have almost universally found positive effects (Wong et al., 2008). A recent example is the age-cutoff RDD study of the program in Boston carried out by Weiland and Yoshikawa (2013) that has received attention for its very positive findings. A disadvantage of the RDD, however, is that it does not allow investigation of pre-k effects after entry into kindergarten because, by then, the control group has also completed pre-k.

These studies demonstrate that state-funded pre-k programs can produce positive effects on various target outcomes prior to kindergarten entry. However, questions about the nature of the effects have been raised. This research has focused on basic pre-reading skills, but the influence of pre-k programs on other outcomes pertinent to children's cognitive and behavioral development, such as complex language skills, mathematics, self-regulation, and social skills, is less clear (Gormley et al., 2005; Jackson et al., 2007). Skills of this latter sort may be more critical for children's long-term performance in school and beyond and thus make better targets for pre-k intervention (Bailey, Duncan, Odgers, & Yu, 2017).

2. Long-term pre-k effects

State investments in pre-k are most often justified by the expectation of long-term effects (e.g., Executive Office of the President of the United States, 2014; Heckman, 2006). This expectation derives mainly from longitudinal research that reported positive outcomes on school completion, employment, marriage stability, criminal behavior, and the like for two model programs – Perry Preschool, mounted in the 1960s, and Abecedarian, begun in the 1970s. Both programs served a small number of children in a single location, and neither has been fully replicated in contemporary publicly funded programs. Indeed, the political feasibility of implementing them at scale is doubtful. These programs would cost more than any state currently allocates – $20,000 per child per year in today's dollars to implement Perry and $16,000–$40,000 for Abecedarian (Minervino & Pianta, 2014).

Attempts to evaluate longer-term effects of state-funded pre-k programs implemented in more recent times have been problematic. Random assignment of children to conditions in which some attend pre-k and others do not, or matching on relevant cognitive, family, and demographic baseline variables, requires that the research sample be identified prior to the beginning of the pre-k year. However, because state pre-k is voluntary, there are few situations in which families intending to enroll their children can be identified in advance and persuaded to participate in random assignment or provide adequate baseline data for matching those who follow through with enrollment and those who do not.

As a result, the largest group of studies of longer-term state pre-k effects compares outcomes for children identified sometime after the pre-k year who did and did not attend pre-k (e.g., Andrews, Jargowsky, & Kuhne, 2012; Gormley, Phillips, & Anderson, 2018; Huang, Invernizzi, & Drake, 2012; Peisner-Feinberg, Mokrova, & Anderson, 2017). These post hoc studies lacking both random assignment and true baseline measures collected prior to the pre-k year are quite vulnerable to selection bias from initial differences on unobserved variables. In short, why did some parents take advantage of a voluntary pre-k program while others did not, and how is that related to family and child characteristics that might influence later outcomes? The demographic variables collected in later years with which these samples are typically matched are unlikely to be sufficient to account for all the relevant differences between children whose parents made and sustained the effort to have them attend pre-k and those who did not.

Another distinct group of studies of longer-term effects of state-funded pre-k programs uses difference-in-difference (DD) methods that examine student outcomes before and after states or counties increased pre-k implementation compared to differences over the same period for areas with no analogous pre-k expansion. The challenge for these studies is to isolate the difference made by variation in pre-k implementation from other influential factors occurring in the same locations over the same period. To do so, they rely on complex statistical models, but those do not always yield robust results. Fitzpatrick (2008), for example, used a DD design to investigate the effects of the Georgia universal pre-k program that grew from participation rates of 14% in 1995 to 55% in 2008. Some analysis models showed positive effects on 4th grade NAEP reading and math scores while others did not. Similar effects that were generally positive, but sensitive to the selection of comparison states, were reported by Cascio and Schanzenbach (2013) for the Georgia and Oklahoma programs. In contrast, DD analyses with extensive data on the More at Four pre-k program in North Carolina showed effects on 3rd grade state achievement scores that were robust to a range of model variations, though the authors acknowledged that the resulting estimates were too large to plausibly represent direct pre-k effects and hypothesized that there must be spillover to nonparticipating children (Ladd, Muschkin, & Dodge, 2014).

The overarching theme in research on long-term effects of state pre-k programs is one of methodological challenge. When dealing with a voluntary program with children's participation always a matter of self-selection by parents, it is difficult for researchers to ensure that they are comparing outcomes for pre-k participants and nonparticipants who are similar in all ways that matter prior to their differential pre-k experience. The result is an uneven and inconclusive research literature. As the experts assembled by the Brookings Institute who recently reviewed virtually all the research on the effects of state pre-k programs reported, "Convincing evidence on the longer-term impacts of scaled-up pre-k programs on academic outcomes and school progress is sparse, precluding broad conclusions" (Phillips et al., 2017, p. 9).

In this context, the Head Start Impact study (Puma, Bell, Cook, & Heid, 2010; Puma et al., 2012) warrants attention. While not a study of state pre-k, it is the only previous randomized study of a public pre-k program. This study began in 2002 with a national sample of 5000 children who applied to 84 programs expected to have more applicants than spaces. Children were randomly selected for offers of admission with those not selected providing the control group. The 4-year-old children admitted to Head Start made greater gains across the pre-k year than nonparticipating children on measures of language and literacy, although not on math. However, by the end of kindergarten the control children had caught up on most achievement outcomes; subsequent positive effects for Head Start participants were found on only one achievement measure at the end of 1st grade and another at the end of 3rd grade. There were no statistically significant effects on social–emotional measures at the end of the pre-k or kindergarten years. A few positive effects appeared in parent reports at the end of the 1st and 3rd grade years, but teacher and child reports in those years showed either null or negative effects.

The positive short-term effects found in the Head Start study are consistent with those found for state pre-k programs. The mixed and null effects found thereafter in this methodologically strong study, however, raise questions about the expectation of substantial long-term benefits that has largely motivated investments in

Please cite this article in press as: Lipsey, M. W., et al. Effects of the Tennessee Prekindergarten Program on children's achievement and behavior through third grade. *Early Childhood Research Quarterly* (2018), https://doi.org/10.1016/j.ecresq.2018.03.005

public pre-k. Nonetheless, the Head Start study is only one randomized control study whereas a body of such research is needed to fully inform policy on public pre-k. The current study contributes a second randomized study to that body of research, one specifically evaluating a state-funded pre-k program.

3. Tennessee Voluntary Pre-k Program

In 1996, the Tennessee legislature funded 10 pilot pre-k programs for children of parents below the federal poverty level (TN Comptroller of the Treasury, 2009). Two years later, the program was expanded and the qualifications were relaxed to include families eligible for the federal free or reduced price lunch (FRPL) programs. By 2008–2009, when the current evaluation began, annual funding had reached $83 million, supporting 938 classrooms and more than 18,000 children, and has remained at about that level ($87 million in 2015–2016; Parker et al., 2016).

Tennessee has funded its pre-k program as a separate enterprise housed in the TN State Department of Education (TNDOE). Local districts apply for funding for pre-k classrooms and receive an amount based on the state's Basic Education Program formula with the remaining costs covered by local funds (Tennessee Alliance for Early Education, 2008). The program is limited to 4-year-old children eligible for kindergarten the following year with families qualifying for FRPL the top priority. A minimum instructional time of 5½ hours per day 5 days a week must be provided in classes of no more than 20 students by a state-licensed teacher endorsed for early childhood education and paid on the same scale as K-12 teachers. A CDA or early childhood associate degree is preferred for educational assistants assigned to each classroom, but not required (TNDOE, 2013). The program must use an approved curriculum from a list provided by TNDOE (TNDOE Office of Early Learning, 2014).

TNDOE attempted to follow the few guidelines available for a quality pre-k program, particularly those from the National Institute of Early Education Research (NIEER, nd). Requirements for programs are laid out in the *Scope of Services* (TNDOE, 2013) and meet 9 of the 10 standards advocated by NIEER. Districts are given latitude within the parameters of those requirements, however, and there is considerable diversity in local implementations. In this regard, the Tennessee Voluntary Pre-K Program (VPK) is not atypical of state pre-k programs generally, operating with some mandated structure based on accepted standards, but neither tightly controlled nor shaped and guided by an overarching vision widely understood and embraced throughout the state. The research on its effectiveness reported here should inform policy makers in states with similar programs, and has already done so for Tennessee policy makers. The guiding questions for this research are as follows.

1. Does participation in VPK improve early literacy, language, and math skills, and classroom behavior by kindergarten entry the following year (immediate effects)?
2. Does participation in VPK provide advantages that carry forward to enhance academic performance in later grades by improving achievement and reducing grade retention, absenteeism, disciplinary infractions, and special education placement (longer-term effects)?
3. Are there demographic subgroups of children who benefit more from participation in VPK?

4. Method

This study is part of a larger VPK evaluation with two major components: a randomized control trial (RCT) implemented in selected oversubscribed sites, and an age-cutoff RDD applied to a probability sample of VPK classrooms across Tennessee. The RDD component was designed to assess variation in classroom characteristics and achievement gains; those results will be reported separately and are not discussed here. However, because the RDD was based on a statewide probability sample, those data were used to create weighting functions for statistical analysis to adjust the RCT effect estimates to represent those expected if they were generalized to the statewide population of participating programs and children.

The RCT reported here was designed to evaluate the effects of VPK participation on a range of educationally relevant outcomes through 3rd grade. The *full randomized sample* of participants in the RCT includes children randomly assigned to receive an offer of admission to VPK or not. These children have been followed in the state education database with attention to attendance, retention in grade, special education designations, disciplinary actions, and state achievement test scores. However, state data provide a limited picture of children's academic performance and classroom behavior during the years prior to 3rd grade when state achievement tests are first administered. The state data were therefore supplemented with data collected by the research team from a subset of children in the full sample with parental consent. These children constitute the *intensive substudy sample* and the data they provided allow the immediate effects of VPK to be assessed. Prior reports have described the components of the overall study and presented findings from earlier stages of the project (Lipsey, Farran, Bilbrey, Hofer, & Dong, 2011; Lipsey, Hofer, Dong, Farran, & Bilbrey, 2013).

4.1. Full randomized sample

In school year 2008–2009, when this study began, VPK had capacity for only 42% of the eligible children in Tennessee (Grehan et al., 2011) and many programs received more applicants than they could accommodate with some children necessarily denied admission. Prior to school years 2009–2010 and 2010–2011, personnel in the TN Office of Early Learning (OEL) surveyed programs about the spaces available in their classrooms and the number of applications they had already received. Based on that information and the experience of the OEL Director with VPK application trends, programs likely to be oversubscribed were identified, informed by OEL that they were eligible for the randomization, and asked if they would participate in the study.

Most of those programs agreed to participate and sent their applicant lists to the research team to be sorted into random order and returned. Program staff were asked to fill their pre-k seats in the order that children appeared on these randomized lists. The procedure was refined for the second cohort to ask program staff to attempt to contact a parent at least three times on different days and times to offer admission. Only if they were unable to contact a parent after these attempts or the parent declined the offer, were they to go to the next child on the list whose parents had not yet been contacted. Once the slots in a given program were filled, the remaining children became a waiting list. If a child offered admission did not show up for the program when school started, the next child on the waiting list was offered that place. Any children not offered admission after that point became eligible for the control group of VPK nonparticipants.

Across two cohorts (school years 2009–2010 and 2010–2011) this procedure generated 150 randomized applicant lists (R-Lists), not all of which were active for the purposes of the study. Active R-Lists involved full VPK classrooms (not blended with children supported by other sources), were actually used to sequence admission offers, and included at least one VPK eligible child who attended a VPK classroom in the R-List program or another VPK program and at least one eligible child who did not attend any VPK classroom. These criteria identified 111 active R-Lists from 79 programs in 29 school districts that included 3131 VPK eli-

Please cite this article in press as: Lipsey, M. W., et al. Effects of the Tennessee Prekindergarten Program on children's achievement and behavior through third grade. *Early Childhood Research Quarterly* (2018), https://doi.org/10.1016/j.ecresq.2018.03.005

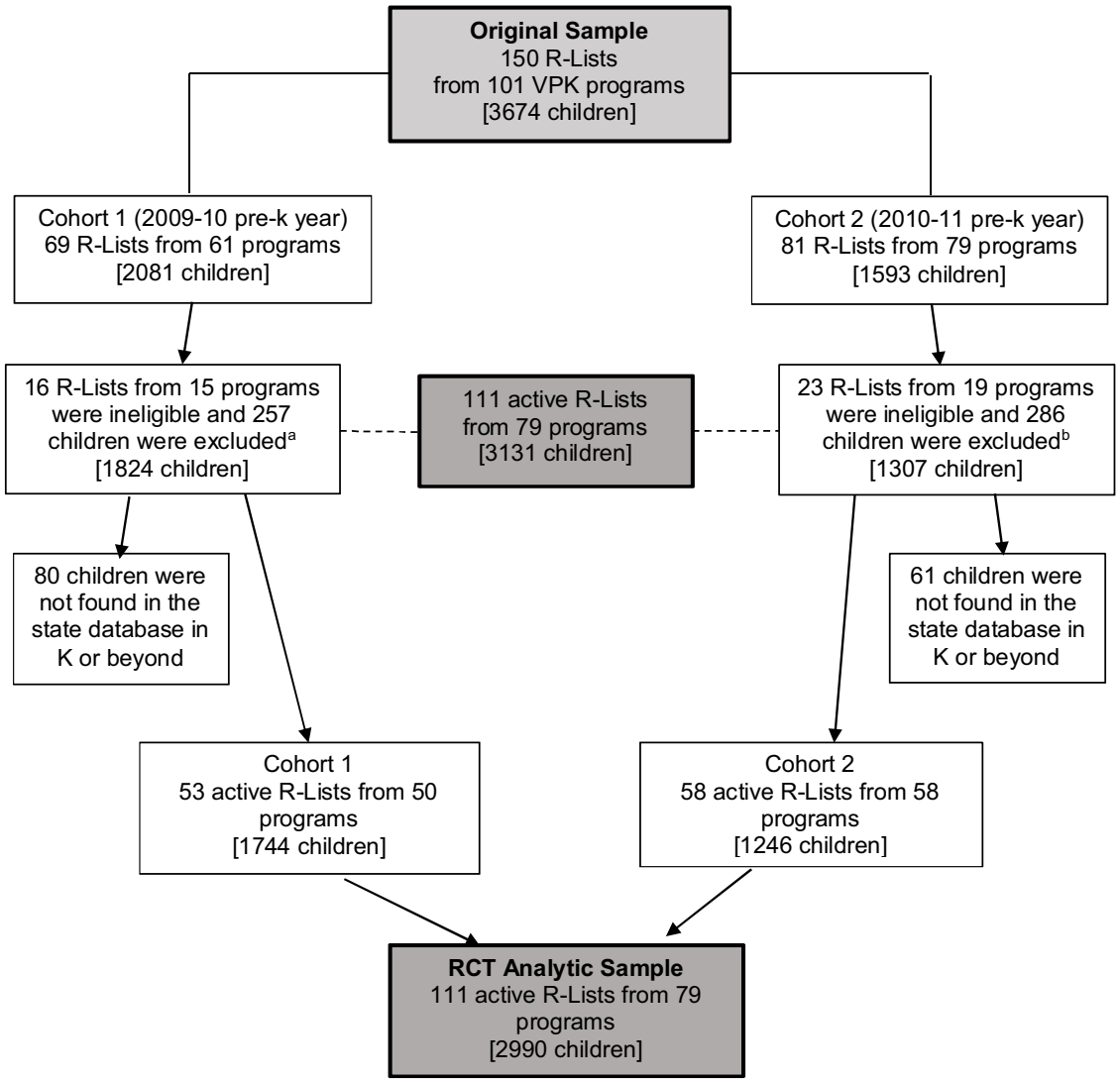

Fig. 1. Construction of the RCT analytic sample.
[a]In Cohort 1, 2 R-Lists from 2 programs with 44 children were lost because the schools did not use the correct random order to admit students. 14 R-Lists from 14 programs with 191 children did not have children in both the T and C conditions, 1 list of which was lost when 1 child who was not eligible for VPK due to age was excluded. An additional 18 children were excluded because they were not eligible for VPK (6 children were in blended classrooms, 1 child was over the income limit, 11 children were too young or too old) but no R-Lists were lost due to these exclusions. 4 children were excluded because they were initially included in a randomized list, but the schools later decided to exclude them from randomization; no R-Lists were lost due to these exclusions.
[b]In Cohort 2, 2 R-Lists from 2 programs with 37 children were lost because the schools did not use the correct random order to admit students. 21 R-Lists from 19 programs with 213 children did not have children in both the T and C conditions. An additional 26 children were excluded because they were not eligible for VPK (12 children were in blended classrooms, 1 child was over the income limit, 13 children were too young or too old) but no R-Lists were lost due to these exclusions. 10 children were excluded because they were initially included in a randomized list, but the schools later decided to exclude them from randomization; no R-Lists were lost due to these exclusions.

gible children. Of those 3131 children, 2990 had a record in the state education database for at least one year of attendance in a Tennessee public school after the pre-k year. Those 2990 children constitute the analytic sample for the RCT (hereafter the *RCT sample*). No evidence that any of the remaining 141 children attended public schools in TN after the pre-k year was found, but 11 of those children had parental consent to participate in the intensive substudy and were retained in that sample (ISS subsample) where they were tracked while homeschooled or enrolled in private schools. Fig. 1 provides details for the construction of the RCT sample.

4.2. Representativeness of the RCT sample

More than 600 programs were funded in each of the years for which the 79 participating programs contributed active R-Lists to the RCT. A natural question is how representative that relatively small group of oversubscribed programs is of VPK programs statewide. Geographically, they were widely distributed across Tennessee with a mix of urban, suburban, and rural locations. A more specific picture was obtained by comparing their characteristics with those of the full population of VPK programs. The characteristics OEL staff identified as most relevant for describing differences among the TN programs were (a) region of the state, (b) urban vs. nonurban location, (c) in a school vs. a partner community agency, (d) an original pilot program vs. one added when VPK went to scale, and (e) whether located in or associated with a high priority school (designated as among the lowest performing in the state). OEL provided descriptive data on those characteristics for the 646 programs funded in 2009–2010. Table 1 shows how the programs in the RCT sample compare on these characteristics with the statewide population of VPK programs.

As the first two panels in Table 1 indicate, the 79 programs in the RCT are distributed across all the categories in this breakdown with proportions similar to the statewide distribution on some of

Table 1
Comparison of the VPK programs contributing to the RCT sample with the statewide population of programs funded in 2009–2010 on key program characteristics.

	Statewide population ($N = 646$)		RCT sample ($N = 79$)		ISS subsample ($N = 58$)	
Program characteristic	Number	Percent	Number	Percent	Number	Percent
Region						
Central west	152	23.5	33	41.8	25	43.1
West	184	28.5	14	17.7	9	15.5
Central east	150	23.2	15	19.0	12	20.7
East	160	24.8	17	21.5	12	20.7
Urbanicity						
Not urban	497	76.9	58	73.4	40	69.0
Urban	149	23.1	21	26.6	18	31.0
Partner site						
Not partner	567	87.8	77	97.5	57	98.3
Partner	79	12.2	2	2.5	1	1.7
Pilot site						
Not pilot	557	86.2	74	93.7	54	93.1
Pilot	89	13.8	5	6.3	4	6.9
Priority						
Not priority	621	96.1	76	96.2	55	94.8
Priority	25	3.9	3	3.8	3	5.2

Notes: Partner sites are those where VPK is provided by a community agency in affiliation with a school district. Pilot sites are those from the initial pilot program that have continued to be active. Priority schools are those officially designated as among the lowest performing in the state.

Table 2
Demographic characteristics of children in the RCT sample and ISS subsample compared to those of VPK participants statewide.

	Statewide sample ($N = 2093$)		RCT sample ($N = 2990$)			ISS subsample ($N = 1076$)		
			Unweighted		Weighted	Unweighted		Weighted
Child characteristic	Number	Mean or percent	Number	Mean or percent	Mean or percent	Number	Mean or percent	Mean or percent
Age (months)	2093	52.7	2990	53.3	52.9	1076	53.2	52.7
Male	1035	49.5	1496	49.4	49.8	512	47.6	49.3
White	1261	60.2	1461	48.9	58.9	605	56.2	61.0
Black	616	29.4	810	27.1	30.3	246	22.9	29.1
Hispanic	196	9.4	673	22.5	9.8	201	18.7	8.5
Non-English	142	6.8	718	24.0	6.8	215	20.0	6.5

Notes: Hispanic refers to ethnicity and children in that category are not also categorized as White to keep the counts independent. Non-English means that English is not the native language.

the characteristics. The largest discrepancy is over-representation in the RCT of programs in the Central West (which includes metropolitan Nashville). It was also possible to compare demographic characteristics of the children in the RCT with those of children in the statewide VPK population. The program characteristics in Table 1 were used as strata in the RDD study mentioned earlier to select a representative sample of VPK programs across the state. Data collected from the children in that RDD sample included demographic characteristics that can be compared to those of the children in the RCT sample. As the first two panels of Table 2 show, there are children in the RCT sample from all the demographic groups found in the statewide population, and with similar proportions on several, but not all, characteristics. The largest discrepancies are an under-representation of White children in the RCT sample, an overrepresentation of Hispanic children, and relatedly, of non-native English speaking children.

In order to extrapolate results of statistical analyses with data from the RCT sample to estimate the analogous results for the statewide population of VPK children, the comparative data in Table 2 were used to create a weighting function for application in those analyses. The child characteristics in Table 2 were used in a logistic regression to predict membership in the RCT sample vs. the statewide probability sample. The predicted values for children in the RCT were thus a kind of propensity score for how similar their characteristics were to those of the children in the population sample. The RCT and population samples were then stratified into bands with similar propensity scores and weights were assigned to the children in the RCT to adjust their proportion in each strata to match the proportion of children from the population in that strata. The third column of the section on the RCT sample in Table 2 shows that applying that weighting function creates a close match between the demographics characteristics of the RCT sample and those of the children in the statewide population. This weighting function is used in analyses reported later to estimate how effects found in the RCT sample are expected to look when generalized to the statewide population of VPK participants.

4.3. Treatment and control groups in the RCT sample

The R-List randomization of the 2990 children in the RCT analytic sample designated 1852 children to receive offers of admission, leaving 1138 who were not to receive such offers. These two groups constitute the intent-to-treat (ITT) treatment and control groups for the RCT. Identification of actual participation in VPK irrespective of the R-List assignment was determined from the state database for attendance during the respective pre-k years for each cohort. Of the 2990 children, 1997 attended VPK for at least one day (mean of 144 days); the remaining 993 children did not. These two groups constitute the treatment-on-treated (TOT) treatment and control groups respectively. Fig. 2 provides details about these comparison groups.

4.4. Intensive substudy subsample

Attempts were made to contact the parents of all 3131 children in the initial RCT sample at the beginning of the school year for each cohort to request consent for annual individual assessments of their children. Though very few explicitly refused, making contact

Please cite this article in press as: Lipsey, M. W., et al. Effects of the Tennessee Prekindergarten Program on children's achievement and behavior through third grade. *Early Childhood Research Quarterly* (2018), https://doi.org/10.1016/j.ecresq.2018.03.005

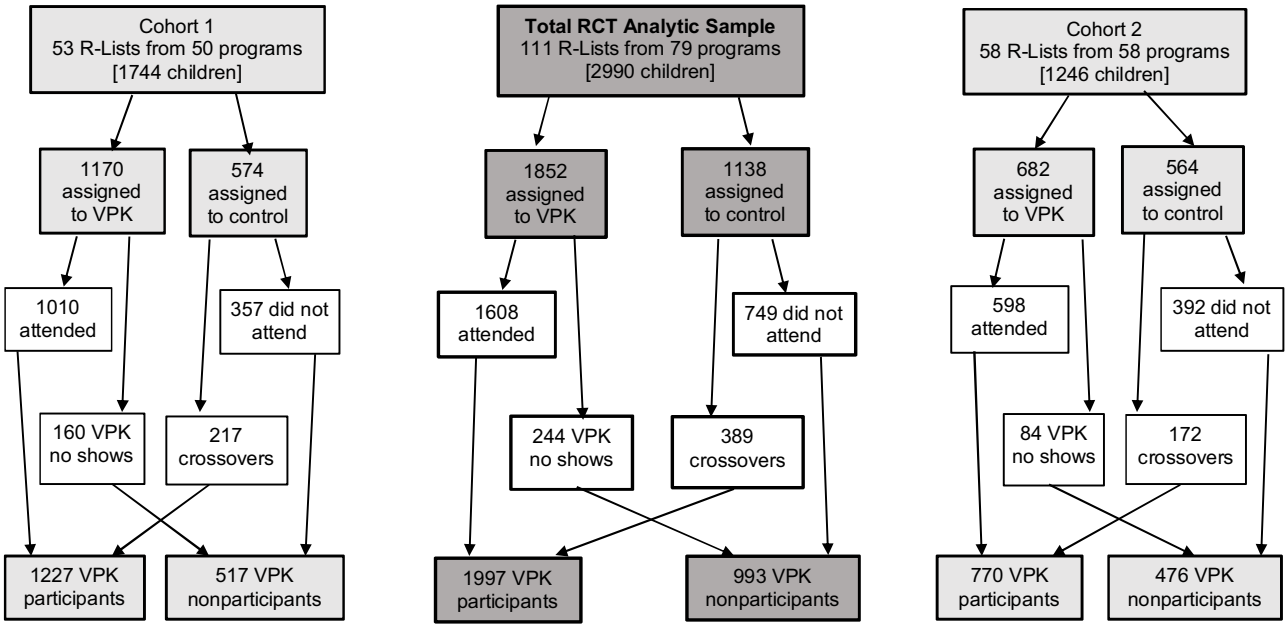

Fig. 2. Composition of the intent-to-treat (ITT) and treatment-on-the-treated (TOT) groups in the RCT sample.

and obtaining responses from parents proved challenging. For the 2009–2010 cohort, TNDOE officials interpreted the confidentiality requirements for FRPL eligible children in a way that only allowed parents to be contacted via a mailing sent centrally from TNDOE. While virtually all the parents who responded consented, most did not respond. Of the 1824 children in that cohort, affirmative parental consent forms were received for only 24.4%.

For the 2010–2011 cohort, arrangements were negotiated to allow parents to be approached about consent as an adjunct to the VPK application process. Participating programs included consent forms with the application paperwork during the registration period, and a member of the research team was available to respond to any questions that arose. This procedure not only facilitated the consent process but had the advantage of requesting parental consent prior to randomization so parents did not yet know if their children would be offered admission. Of the initial 1307 children in that cohort, affirmative consents were received for 67.8%. Again, most of the remainder simply did not respond; very few actively declined to consent.

These procedures yielded 1331 consented children, not all of whom were eligible for the intensive substudy sample. As with the full RCT sample, this subsample was restricted to children age-eligible for kindergarten the next year, income-eligible for FRPL, not in a blended classroom, and who had not applied only to receive out-of-classroom special education services. The sample was further restricted to children on R-Lists with consented children in both the treatment and control groups and who had at least one posttest measure at the end of the pre-k year. These restrictions left an analytic sample (hereafter the Intensive SubStudy subsample; i.e., *ISS subsample*) of 1076 children on 76 R-Lists from 58 VPK programs in 21 school districts across the state. Those 76 R-Lists included 2086 eligible children with the 1076 in the ISS subsample thus representing a 51.6% participation rate from those R-Lists. Fig. 3 reports details for the consent rates and construction of the ISS analytic subsample.

4.5. Representativeness of the ISS subsample

The ISS subsample is a subset of the RCT sample except for the 11 children with parental consent for data collection while home-schooled or enrolled in private school, but with no data past the pre-k year in the state database. Relevant questions are how similar the ISS subsample is to the full RCT sample and, especially, how similar it is to the statewide VPK population. The third panels of Tables 1 and 2 address these questions and show that the characteristics of the programs and children contributing to the ISS subsample are substantially similar to those of the RCT sample and the statewide population.

While Table 2 compares the ISS subsample to VPK participants statewide on demographic characteristics, a more detailed comparison is possible for that subsample. The data collected on the statewide RDD sample included assessments at the beginning of the pre-k year on the same Woodcock–Johnson III (WJ) achievement measures administered at the beginning of the pre-k year to the ISS subsample (described in more detail later). Table 3 shows that the ISS subsample has baseline WJ achievement scores similar to those of the statewide sample. To allow a closer match, a weighting function was created for the ISS subsample using the procedure described above for the RCT sample. The results of applying that weighting function to the baseline characteristics of the children in the ISS subsample are shown in the far right columns of Tables 2 and 3. That weighting function was also applied in analyses reported later to extrapolate findings from the ISS subsample to the statewide population.

4.6. Treatment and control groups in the ISS subsample

Of the 1076 children in the ISS subsample, the R-Lists designated 697 to receive offers of VPK admission while the remaining 379 were not to receive offers. These groups thus constitute the intent-to-treat (ITT) treatment and control groups for analysis with the ISS subsample. According to records in the state database, 780 of the 1076 children attended VPK for at least one day (mean 146 days) with no attendance recorded for the remaining 296 children. These two groups constitute the treatment-on-treated (TOT) treatment and control groups for analysis. Fig. 4 provides more detail about the composition of these comparison groups in the ISS subsample.

For children who did not attend VPK, parent interviews identified the alternative arrangements made for their children during the pre-k year. A majority of parents reported that their VPK eligible

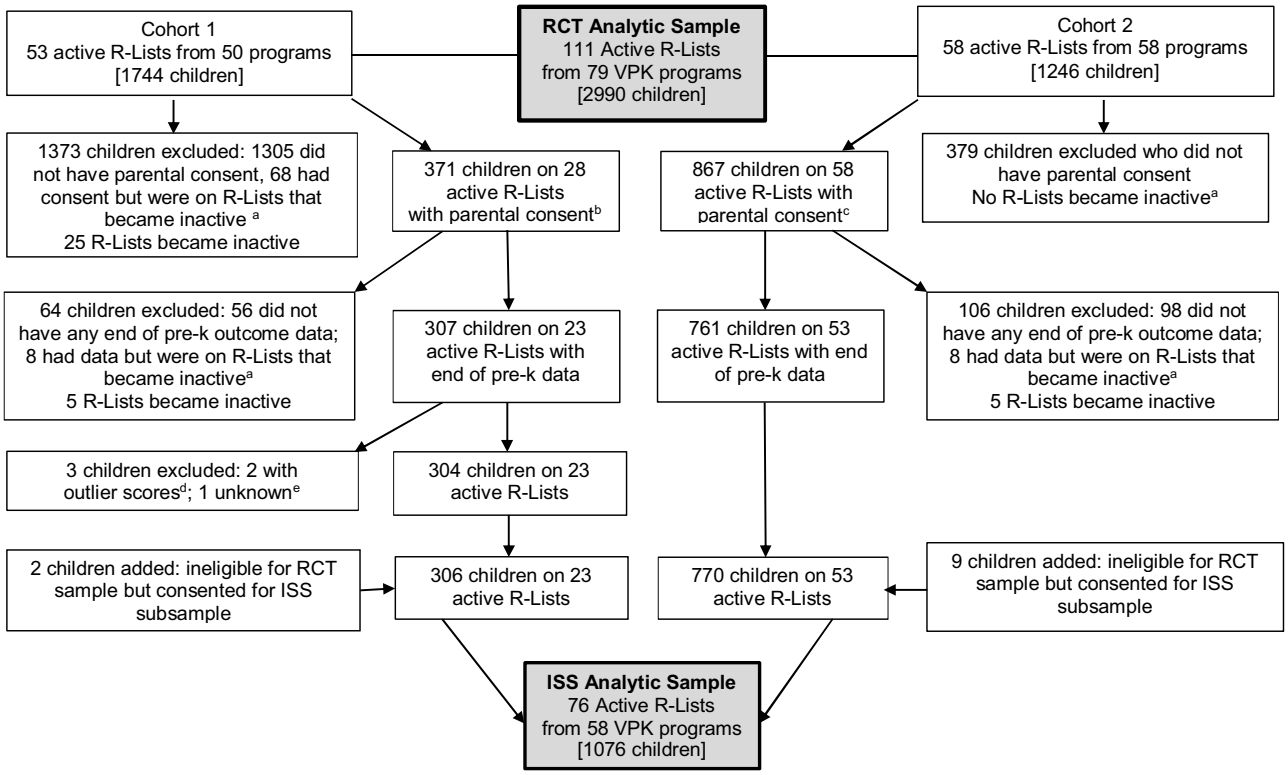

Fig. 3. Construction of the ISS analytic subsample.
[a]R-Lists became inactive for use in the ISS subsample when they no longer had at least one consented child who attended VPK and one who did not attend.
[b]Of the 337 children in Cohort 1 ineligible for the RCT analytic sample, parental consent for participation in the ISS subsample was obtained for 11 children.
[c]Of the 347 children in Cohort 2 ineligible for the RCT analytic sample, parental consent for participation in the ISS subsample was obtained for 72 children.
[d]Distributions of propensity scores created from baseline variables were compared for the Cohort 1 VPK participant and nonparticipant groups in a preliminary screening. Two children with outlier scores at opposite extremes of those distributions were trimmed from the sample on the basis of that screening.
[e]Child was initially identified as ineligible for the ISS subsample, but that may have been an error. There is insufficient information to resolve this ambiguity.

Table 3
Scores on Woodcock–Johnson achievement measures at the beginning of the pre-k year for children in the ISS subsample compared to those of VPK participants statewide.

	Statewide sample ($N = 2093$)		ISS subsample ($N = 1076$)					
			Unweighted			Weighted		
Achievement measure[a]	Mean	SD	Mean	SD	Diff in SD units[b]	Mean	SD	Diff in SD units[b]
WJ Composite 6	393.9	15.5	394.3	18.1	−0.022	393.8	16.9	0.010
Letter-Word ID	317.3	23.9	318.6	27.1	−0.054	317.3	26.5	−0.002
Spelling	349.1	22.2	351.1	28.4	−0.091	348.7	28.5	0.017
Oral Comprehension	443.4	15.0	443.3	16.2	0.005	443.1	15.2	0.014
Picture Vocabulary	457.7	18.6	454.6	23.3	0.171	458.1	17.7	−0.019
Applied Problems	389.9	26.9	390.6	27.8	−0.027	389.1	26.7	0.029
Quantitative Concepts	406.3	12.9	407.6	14.0	−0.097	406.2	13.5	0.009

[a] Woodcock–Johnson III achievement measures; the longitudinally scaled W-scores are reported. WJ Composite 6 is the mean of the W-scores across the six content-specific scales included in this table.
[b] Based on the standard deviations of the statewide sample.

child did not attend any center-based preschool program. Overall, 63% received home-based care by a parent, relative, or other person; 13% attended Head Start or what parents described as a public pre-k program; 16% were in private center-based childcare; 5% had some combination of Head Start and private childcare; and childcare for 3% was not reported.

4.7. Data collection

Data were drawn from the state database for the 2990 children in the RCT analytic sample for each year through the 3rd grade year. Those data included descriptive characteristics such as birthdate, gender, race/ethnicity, and native language as well as outcome variables for attendance, retention in grade, disciplinary actions, special education designations, and scores on the state achievement tests. Some children were not enrolled in a Tennessee public school in some years and did not appear in the state records those years. Records were found for 98.5% of the RCT sample in the kindergarten year, 97.8% in the 1st grade year, 95.8% in the 2nd grade year, and 93.7% in the 3rd grade year. Note that throughout this report phrases like "1st grade year" refer to the year in which children were expected to be in that grade according to the normal sequence, including those retained in a prior grade and thus not actually in the indicated grade.

Children in the ISS subsample who attended Tennessee public schools after the pre-k year (1065 of the 1076) also had data

Please cite this article in press as: Lipsey, M. W., et al. Effects of the Tennessee Prekindergarten Program on children's achievement and behavior through third grade. *Early Childhood Research Quarterly* (2018), https://doi.org/10.1016/j.ecresq.2018.03.005

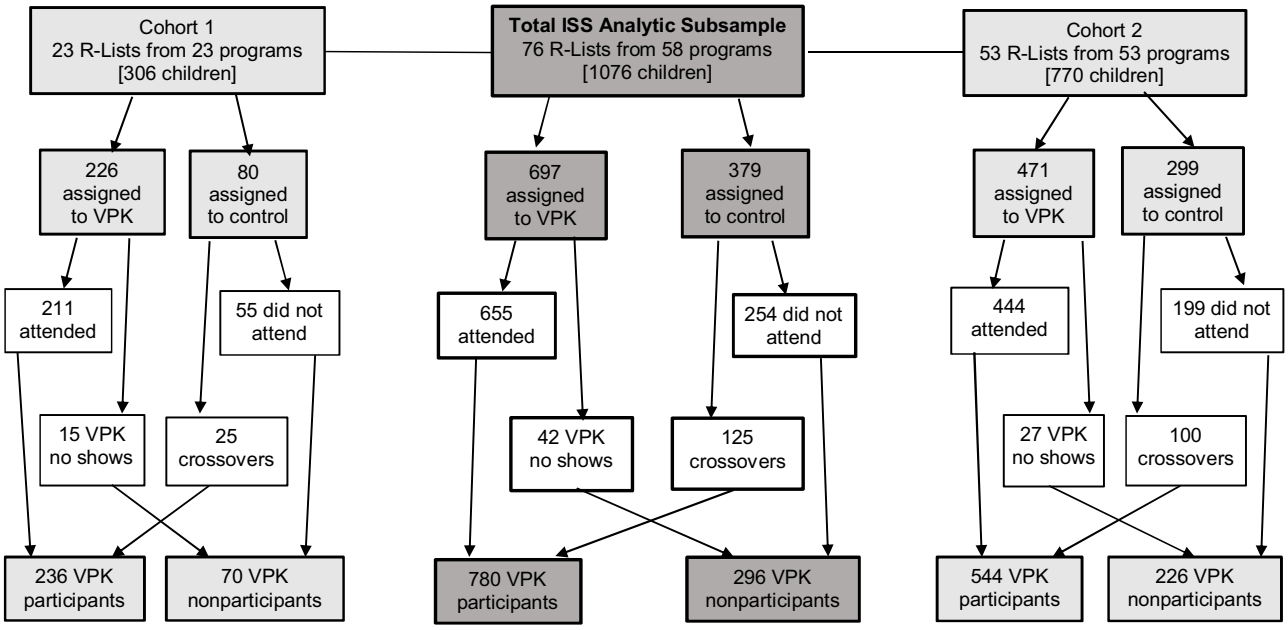

Fig. 4. Composition of the intent-to-treat (ITT) and treatment-on-the-treated (TOT) groups in the ISS subsample.

from the state database. In addition, the children in that subsample were individually assessed by research staff in the fall and spring of their pre-k year (including the 11 children home schooled or attending private schools). VPK participants were assessed at their program sites and nonparticipants were assessed at locations convenient for the parents (e.g., Head Start centers, libraries, parks, and homes). Children in both groups were assessed in the spring of each subsequent year through the 3rd grade year. Assessments were completed on 100% of the ISS subsample at the end of the pre-k year, 97.2% at end of kindergarten, 95.6% at end of 1st, 94.4% at end of 2nd, and 92.3% at end of 3rd.

Early in the kindergarten year and in the spring of the 1st through 3rd grade years, children's classroom behaviors were rated by their teachers. Ratings by kindergarten teachers near the beginning of the kindergarten year were treated as pre-k outcomes. The target time for those ratings was 6 weeks (42 days) into the kindergarten year – long enough for teachers to become familiar with the children but before extensive exposure to kindergarten instruction. Not all teachers were responsive to that request – on average, ratings were received 52 days into the kindergarten year for VPK participants and 51 days for control children (standard deviation of 41 days). The completion rates for teacher ratings ranged from 89.5% at the beginning of kindergarten to 86.7% at the end of the 3rd grade year.

4.8. Measures for the intensive substudy (ISS) subsample

4.8.1. Parent questionnaire

During the pre-k year, parents of consented children were interviewed via telephone about their education and employment and that of their spouse/partner, the home language and literacy environment, and childcare arrangements if their child was not in VPK. When needed, these interviews were conducted by Spanish-speaking interviewers.

4.8.2. Direct assessments

Children's academic achievement was assessed with a selection of scales from the Woodcock–Johnson (WJ) III Achievement Battery (Woodcock, McGrew, & Mather, 2001) administered in English. At the beginning and end of the pre-k year, these included two measures of literacy skills (Letter–Word Identification and Spelling), two measures of language skills (Oral Comprehension and Picture Vocabulary), and two measures of math skills (Applied Problems and Quantitative Concepts). At the end of the kindergarten year and through the 3rd grade year two other scales were added: a reading measure (Passage Comprehension) and another math measure (Calculation). Analysis was conducted with the longitudinally scaled W-scores from these measures, which are comparable from year to year in ways that index the change that has taken place over successive waves of measurement. The scores on these different WJ scales were moderately to highly intercorrelated. To provide summary achievement indices, composite scores were created as the mean of the W-scores across the individual scales, one that combined the original six scales administered from the beginning of pre-k (WJ Composite 6) and another that combined those six subscales with the two first administered at the end of the kindergarten year (WJ Composite 8).

4.8.3. Teacher ratings

Two teacher rating instruments were completed by kindergarten, 1st, 2nd, and 3rd grade teachers. The Cooper–Farran Behavioral Rating Scales (Cooper & Farran, 1991) ask teachers to rate each child's work-related and interpersonal skills. *Work-Related Skills* assesses ability to work independently, listen to the teacher, remember and comply with instructions, complete tasks, and otherwise engage appropriately in classroom activities. This scale consists of 16 7-point behaviorally anchored ratings with the mean as an overall total score (Cronbach alpha reliability coefficients ranged from .94 to .95 across the waves of data collection). *Interpersonal Skills* assesses social interactions with peers including behavior in group activities, play, and outdoor games; expression of feelings and ideas; and response to others' mistakes or misfortunes. It consists of 21 7-point behaviorally anchored ratings with the mean providing an overall total score (Cronbach alphas from .93 to .94 across waves).

The second measure, the Academic Classroom and Behavior Record (ACBR; Farran, Bilbrey, & Lipsey, 2003), consists of four scales. *Preparedness for Grade Level Work* asks how prepared the child is in math, literacy/language, and social behavior. It includes three 7-point behaviorally anchored ratings with the mean as

Please cite this article in press as: Lipsey, M. W., et al. Effects of the Tennessee Prekindergarten Program on children's achievement and behavior through third grade. *Early Childhood Research Quarterly* (2018), https://doi.org/10.1016/j.ecresq.2018.03.005

Table 4
Number and proportion of children randomly assigned to VPK participation or nonparticipation who did and did not comply with that assignment.

	RCT sample ($N = 2990$)		ISS subsample ($N = 1076$)	
Random assignment	Participants number (proportion)	Nonparticipants number (proportion)	Participants number (proportion)	Nonparticipants number (proportion)
VPK (treatment)	1608 (.87)	244 (.13)	655 (.94)	42 (.06)
No VPK (control)	389 (.34)	749 (.66)	125 (.33)	254 (.67)
ITT to TOT multiplier	1/(.8683−.3418) = 1.8993		1/(.9397−.3298) = 1.6396	

Notes: Data for full samples; assumes imputation of missing outcome values. Participants/Nonparticipants refers to those who did or did not attend VPK irrespective of the random assignment. ITT to TOT multiplier is the inverse of the proportion assigned to treatment that participated minus the proportion assigned to control that participated.

the total score (Cronbach alphas from .85 to .87 across the measurement waves). *Peer Relations* includes two 7-point behaviorally anchored ratings for whether other children like the child and how many close friends the child has; the mean is the total score (Cronbach alphas of .75 to .80 across waves). *Behavior Problems* asks whether a child has shown any of nine problem behaviors including explosive or overactive behavior, attention problems, physical or relational aggression, and social withdrawal or anxiety. A total score is computed as 0 or 1 to indicate whether no problems are identified vs. one or more (Cronbach alphas of .63–.70 across waves). *Feelings About School* consists of six 3-point ratings for the child's liking or disliking school, enjoying and engaging in classroom activities, and seeming happy at school. Ratings are skewed toward positive responses so the total score is computed as the mean with the lowest rating on any item scored 1 and either of the higher ratings scored 2 (Cronbach alphas of .80–.85 across waves).

4.9. Analysis

4.9.1. Missing data

Missing value rates for the RCT sample ranged from 0% to 6.3% except for 3rd grade state achievement scores with 18.4% missing, mostly due to children retained who had not yet reached 3rd grade. Missing values for the ISS subsample ranged from 0% to 5.6% for parent interview and child assessment data and from 10.5% to 13.3% for teacher ratings. To retain the full samples in all analyses given these modest missing data rates, multiple imputation was done separately in the RCT and ISS samples and, in each, separately for VPK participants and nonparticipants using the Mistler (2013) procedure for multilevel data (children nested in R-Lists). Fifty imputed files were produced with the results of analysis of each pooled to include the uncertainty associated with the imputations in the standard error estimates. Analyses using only observed outcome data were conducted in parallel to identify any sensitivity to the imputations in the statistical conclusions.

4.9.2. Analysis models

All analyses used hierarchical linear models with children nested in R-Lists that were nested in school districts. The primary analyses were intent-to-treat (ITT) comparisons with outcomes for children randomly assigned to receive an offer of VPK admission compared with outcomes for children assigned to receive no offer. A standard set of covariates (described later) was included in each analysis to adjust for baseline differences, improve statistical power, and provide a basis for moderator analysis. The results of each ITT analysis were then used to estimate treatment-on-the treated (TOT) effects for children who actually attended VPK compared to children who did not attend. The TOT estimates were generated from two-stage least-squares (2SLS) regressions with R-List random assignment as an instrumental variable (Angrist, Imbens, & Rubin, 1996; Angrist, 2006).

Overall, 79% of the RCT sample and 84% of the ISS subsample complied with the random assignment. The 2SLS estimates for TOT effects were obtained by rescaling the ITT effect estimates with a multiplier based on the proportion (p_{tp}) of children assigned to VPK who actually participated (treatment compliers) and the proportion (p_{cp}) assigned to the control condition who nonetheless also participated (crossovers) that is defined as the inverse of the $p_{tp} - p_{cp}$ difference (reported in Table 4 for the RCT and ISS samples). The rationale for this approach to estimating TOT effects can be found in Puma et al. (2010, pp. 5–34 to 5–53) and Gennetian, Morris, Bos, and Bloom (2005). A notable feature of this procedure is that it rescales the standard errors of the ITT estimates with the same multiplier so that the statistical significance of the TOT effect estimates is the same as that for the ITT estimates.

4.9.3. Baseline equivalence for the RCT sample

Baseline variables for the RCT are limited to the few static demographics available in the state database: age, gender, race/ethnicity, and native language. Table 5 shows the differences between the ITT treatment and control groups on these variables. The differences were tested in multilevel OLS regression models with ITT condition as the only predictor and none was statistically significant. For binary variables, more technically appropriate logistic regression analyses were also conducted to check that the results were similar (Cleary & Angel, 1984), and those also showed no significant differences. There were no missing values on these variables and all but the language differentiation for Hispanic children were included as covariates in the outcome analyses for the RCT reported later.

Further analysis explored the baseline equivalence on these variables for each of the two cohorts of children combined in the full RCT sample. The analyses reported in Table 5 were repeated with addition of dummy codes for cohort and cohort by treatment condition interaction terms that tested for cohort differences on the treatment-control equivalence of the baseline variables. Neither the cohort main effect nor the interaction was statistically significant for any baseline variable. Additionally, all analyses of VPK effects for the RCT sample reported later were repeated with cohort and cohort by treatment condition interaction terms added. With only a few scattered exceptions, these interaction tests showed no statistically significant differential treatment effects for the cohorts so VPK effect estimates are not reported separately by cohort.

Because the ISS subsample is the consented subset of the RCT sample (except for 11 children), the similarity of the ISS subsample to the remaining children not in that subsample on the RCT baseline variables can also be examined. Each analysis reported in Table 5 was repeated with addition of a dummy code for ISS membership and its interaction with the ITT treatment condition. Neither the main effect for ISS subsample membership nor the interaction was statistically significant at $p < .05$ for any baseline variable. Two effects were significant at the $p < .10$ level, one indicating that the proportion of White children in the ISS subsample was somewhat larger than the proportion among those not in that subsample (by about five percentage points), the other for an interaction term that

Please cite this article in press as: Lipsey, M. W., et al. Effects of the Tennessee Prekindergarten Program on children's achievement and behavior through third grade. *Early Childhood Research Quarterly* (2018), https://doi.org/10.1016/j.ecresq.2018.03.005

Table 5
Intent-to-treat (ITT) treatment-control comparison on baseline variables for the RCT sample.

Variable	Treatment group mean[a]	Control group mean[a]	Pooled SD[b]	Coefficient[c] for T-C difference	Effect size[d]	*p*-value
Age (months)	53.2	53.3	3.47	−.087	−.025	.507
Gender (male)	.495	.489	.500	.006	.012	.752
White	.674	.684	.505	−.010	−.019	.578
Black	.196	.195	.451	.001	.003	.941
Hispanic	.140	.132	.411	.007	.018	.639
Hispanic, native English	.021	.029	.158	−.009	−.054	.158
Hispanic, not English	.122	.110	.393	.012	.030	.421
Not native English	.138	.131	.414	.008	.019	.617
	N = 1852	*N* = 1138				

$^{*}p < .05$, $^{\dagger}p < .10$.
[a] Estimated marginal means from multilevel analysis models; except for age, these can be read as proportions. There were no missing values on these baseline variables so no imputed values were used in these analyses.
[b] Pooled treatment and control group standard deviations.
[c] Coefficients for the ITT treatment-control difference from multilevel models with children nested in R-Lists, R-Lists nested in districts, with ITT condition as the only predictor.
[d] Effect size: coefficient for the treatment-control difference divided by the pooled standard deviation.

Table 6
Intent-to-treat (ITT) treatment-control comparison on baseline variables for the ISS subsample.

Variable	Treatment group mean[a]	Control group mean[a]	Pooled SD[b]	Coefficient[c] for T-C difference	Effect size[d]	*p*-value
Age (months)	53.1	53.4	3.43	−.299	−.087	.171
Gender (male)	.468	.483	.500	−.015	−.031	.638
White	.646	.666	.505	−.020	−.039	.505
Black	.236	.216	.432	.020	.047	.445
Hispanic	.140	.155	.385	−.016	−.040	.527
Hispanic, native English	.027	.038	.175	−.011	−.063	.341
Hispanic, not English	.114	.123	.358	−.009	−.025	.701
Not native English	.136	.169	.389	−.034	−.087	.172
Mother's education	2.13	2.11	.743	.014	.019	.786
No. of working parents	1.25	1.24	.635	.008	.013	.846
Library card	.972	.886	.846	.086	.102	.142
No. of newspapers	.382	.331	.775	.051	.066	.328
No. of magazines	.291	.267	.515	.025	.048	.479
Lag to pretest (days)	41.7	52.5	26.96	**−10.77***	**−.322**	**.000**
WJ Composite 6	394.5	395.2	18.3	−.666	−.036	.583
WJ Letter-Word ID	318.4	317.7	27.4	.722	.026	.695
WJ Spelling	349.4	352.2	28.7	−2.84	−.099	.140
WJ Oral Comp	444.2	443.7	16.3	.419	.026	.704
WJ Picture Vocab	456.8	455.6	23.1	1.292	.056	.396
WJ Applied Problems	391.7	392.5	28.0	−.812	−.029	.666
WJ Quant Concepts	406.8	408.9	14.2	**−2.11***	**−.149**	**.027**
	N = 697	*N* = 379				

$^{*}p < .05$, $^{\dagger}p < .10$ for coefficients; significant coefficients and related estimates are also bolded.
Notes: Mother's education: 1 = below high school; 2 = completed high school; 3 = some postsecondary; 4 = bachelors' degree or higher. Library card: 1 = yes for at least one parent, 0 = no. No. of newspapers and magazines counts number of household subscriptions. Lag to pretest is counted from September 1, the start date or very close for VPK programs, to the date of WJ assessments. WJ Composite 6 is the mean of the W-scores across the six content-specific scales included in this table.
[a] Estimated marginal means from the multilevel analysis model.
[b] Pooled treatment and control group standard deviations.
[c] Coefficients for the ITT treatment-control differences from a multilevel model with multiple imputation for missing values and children nested in R-Lists, R-Lists nested in districts, with ITT condition as the only predictor.
[d] Effect size: coefficient for the treatment-control difference divided by the pooled standard deviation.

identified more non-native English speakers in the control group than the treatment group for the ISS subsample, with the reverse pattern for the non-ISS subsample (three percentage point differences in each case). Within each subgroup, however, the baseline treatment-control differences were not significant.

Overall, these results show substantial similarity between children included and not included in the ISS subsample, but also raise the possibility of differential effects for these groups on the outcomes examined later. Accordingly, all analyses of VPK effects with the RCT sample were repeated with a test of the ITT condition by ISS membership interaction added. These many interaction tests identified only a few statistically significant differential VPK effects between the ISS subsample and the remaining portion of the RCT sample. This finding has important implications. In particular, it indicates that VPK effects that can only be examined with the ISS subsample can be expected to generalize reasonably well to the full RCT sample.

4.9.4. *Baseline equivalence for the ISS subsample*

An extensive set of baseline variables is available for the ISS subsample including pretests on the WJ achievement scales as well as family information from the parent interviews. Table 6 compares the ITT treatment and control groups in the ISS subsample on a differentiated set of demographic variables, the WJ pretests, and key variables from the parent interviews. Multilevel tests confirmed the substantial similarity of the groups on all but two of these variables and, except for those, it is notable that the effect sizes for the differences are modest, well under the Imbens and Rubin (2015, p. 277) suggested standard of .25 for baseline differences too large to adjust with covariates.

The most striking exception to the similarity of the ISS treatment and control groups at baseline relates to the time lag between the start of VPK classes and administration of the WJ pretests. Start times across sites clustered closely around September 1, which was used as the date for computing time lags. The pretest assessment

Please cite this article in press as: Lipsey, M. W., et al. Effects of the Tennessee Prekindergarten Program on children's achievement and behavior through third grade. *Early Childhood Research Quarterly* (2018), https://doi.org/10.1016/j.ecresq.2018.03.005

was an average of 45.5 days after that date overall but, as shown in Table 6, it was about 11 days later for the control children than the VPK treatment children. The general equivalence between these groups on the WJ pretests gives no indication that this timing difference affected their scores. Nonetheless, to statistically adjust for any influence of pretest timing difference or any related timing misalignment between the ISS treatment and control groups, pretest time lag and age at time of assessment were included as covariates in all analyses of WJ achievement outcomes. Also included as covariates in analyses with the ISS sample were the WJ Composite 6 pretest to represent initial achievement performance, the Quantitative Concepts pretest (the one WJ scale with a significant pretest difference), and age, gender, race/ethnicity, native English, and mother's education. Despite the few significant differences on the baseline variables, inclusion of these covariates in analyses of VPK effects further ensures the equivalence of the ISS subsample treatment and control groups and improves the statistical power of the analyses for this smaller sample.

5. Results

The research questions that guided this study asked about the effects of VPK on children's cognitive skills and classroom behavior by kindergarten entry, whether VPK effects carry forward to enhance performance in later grades, and whether the effects are greater for some demographic subgroups of children than others. The findings presented in this results section address each of these questions for each outcome domain for which data are available. Full details for any of the analyses reported in this article and the associated sensitivity tests for the analysis models are available from the corresponding author.

5.1. Academic performance

5.1.1. Achievement

For school systems, a major aspiration is that VPK improve the performance of participating children on the state achievement tests first administered in 3rd grade. However, there is an interval of nearly four years between the end of pre-k and 3rd grade during which state data provide no direct information about children's achievement. The primary rationale for the ISS subsample, with its annual assessments on WJ achievement measures, was to shed light on children's performance during that interval.

5.1.2. Woodcock–Johnson achievement measures (ISS)

Table 7 reports the VPK ITT and TOT effects for the WJ achievement measures administered to the ISS subsample from pre-k through the 3rd grade year. The composite measures provide the best overall summary, but the results for the individual measures that contribute to those composites are also presented. Table 7 shows rather consistent findings across the achievement measures. VPK effects are positive and statistically significant at the end of the pre-k year with effect sizes for TOT impacts large enough to be educationally meaningful, e.g., by the .25 threshold used by the U.S. Department of Education What Works Clearinghouse. This is not only true for the composite measure, but for each of the scales represented in that composite with the exception of Oral Comprehension.

However, in later grades the performance of the children in the control conditions converged with that of children in the VPK conditions; that is, the control children caught up to the VPK children. By the end of kindergarten, there were no longer any statistically significant differences on most of the achievement measures (Picture Vocabulary is the only exception, and that difference is no longer significant by the end of the 1st grade year). After the kindergarten year, most of the effect estimates are negative, though short of statistical significance, indicating that the control children outperformed the children in the VPK treatment conditions.

The analyses that generated the results in Table 7 used multiple imputation for missing values; parallel analyses using only observed values produced similar results for statistical conclusions and effect size magnitude. In particular, VPK effects were positive and statistically significant at the end of pre-k for all the achievement measures except Oral Comprehension, with null or negative findings thereafter except for Picture Vocabulary in kindergarten.

To extrapolate the VPK findings from the ISS subsample to the statewide population of VPK participants, the analyses reported in Table 7 were repeated with application of the weights described earlier that create a close match between the ISS subsample and the statewide population on demographic characteristics and baseline WJ scores. The results in the last two columns of Table 7 reveal a pattern of TOT effect estimates for the generalization to the statewide population that is quite similar to the pattern in the ISS subsample. The statewide estimates at the end of pre-k are positive but smaller than those for the ISS subsample with those statistically significant in the ISS subsample remaining so in the statewide estimates except for Applied Problems. Moreover, after pre-k there are negative trends in the statewide estimates that largely parallel those in the ISS subsample, though only a few reached statistical significance.

5.1.3. Teacher ratings of preparedness for grade (ISS)

The data for the ISS subsample include teacher ratings of how prepared children were for grade-level work related to language/literacy, math, and social behavior. Kindergarten teachers made their ratings near the beginning of the school year; 1st, 2nd, and 3rd grade teachers made their ratings near the end of the respective years. The ITT and TOT effects on those ratings (Table 8) showed a pattern similar to that found for the WJ achievement measures (Table 7). Early in the kindergarten year, teachers reported that VPK participants were better prepared than the control group ($p < .10$). However, in the later grades the direction of the effect was reversed with the control group receiving higher ratings, although none of those differences reached statistical significance. Moreover, the weighted analysis that generalized the ISS subsample results to the statewide VPK population showed the same pattern, giving no indication that this finding was distinctive to the ISS subsample.

5.1.4. State achievement tests (RCT)

The Tennessee Comprehensive Assessment Program (TCAP) in place during this study requires annual testing on reading/language arts, mathematics, and science beginning in 3rd grade. The 3rd grade scores on those tests were available for most of the children in the RCT sample although some were retained in a prior grade and thus had not yet reached 3rd grade. However, as reported later, there were no statistically significant differences in retention between the VPK treatment and control groups by the 3rd grade year that would bias the comparisons of 3rd grade TCAP scores for children who did take the tests. Furthermore, our primary analysis models used multiple imputation for missing data, which includes the missing TCAP scores for retained children. Because imputation for this distinctive situation is more questionable than when missingness is not so heavily influenced by a single factor, the results of parallel analyses with only observed values are also reported.

The first panel of Table 9 presents the results with multiple imputation of missing scores; the second panel shows the results with only observed values. The third panel shows the results with multiple imputation and the weighting function that matches the RCT sample to the demographic profile of the statewide VPK population. All these analyses show the same pattern of VPK effects on the state achievement tests. As indicated by the negative effect coefficients and effect sizes, the control children outperformed the

Please cite this article in press as: Lipsey, M. W., et al. Effects of the Tennessee Prekindergarten Program on children's achievement and behavior through third grade. *Early Childhood Research Quarterly* (2018), https://doi.org/10.1016/j.ecresq.2018.03.005

Table 7
Intent-to-treat (ITT) and treatment-on-treated (TOT) impact estimates for the WJ achievement measures (ISS subsample).

	ITT						TOT				TOT weighted	
WJ scale & year	Treatment group mean[a]	Control group mean[a]	Pooled SD[b]	Coefficient[c] for T-C difference	Effect size[d]	*p*-value[e]	Treatment group mean[a]	Control group mean[a]	Coefficient[c] for T-C difference	Effect size[d]	Coefficient[c] for T-C difference	Effect size[d]
WJ Composite 6												
Pre-k	411.1	407.1	16.80	**4.00***	**.236**	**.000**	412.4	405.8	**6.57***	**.395**	**4.30***	**.270**
K	441.0	440.7	13.88	0.27	.019	.676	441.1	440.7	0.44	.032	0.52	.038
1st	464.1	464.3	14.33	−0.21	−.015	.777	464.0	464.4	−0.34	−.024	−0.58	−.040
2nd	478.0	479.1	13.95	−1.12	−.080	.130	477.6	479.5	−1.84	−.132	−1.85	−.127
3rd	489.4	490.4	13.93	−0.98	−.070	.204	489.1	490.7	−1.60	−.115	−1.81	−.125
WJ Composite 8												
Pre-k	–	–	–	–	–	–	–	–	–	–	–	–
K	437.8	438.0	14.06	−0.15	−.011	.818	437.8	438.0	−0.25	−.018	0.03	.002
1st	463.1	463.4	14.24	−0.32	−.022	.682	463.0	463.5	−0.52	−.036	−0.56	−.038
2nd	477.4	478.3	13.32	−0.89	−.067	.210	477.1	478.5	−1.47	−.110	−1.40	−.101
3rd	488.8	489.6	13.45	−0.85	−.063	.259	488.5	489.9	−1.39	−.103	−1.52	−.108
Letter–word ID												
Pre-k	344.0	336.4	26.94	**7.61***	**.279**	**.000**	346.4	333.9	**12.48***	**.470**	**10.60***	**.400**
K	398.5	399.5	27.34	−1.07	−.039	.480	398.1	399.9	−1.76	−.064	−0.47	−.017
1st	446.3	447.5	29.47	−1.15	−.039	.519	446.0	447.8	−1.89	−.064	−1.66	−.053
2nd	470.4	472.3	25.88	−1.96	−.076	.209	469.7	473.0	−3.22	−.124	−2.95	−.108
3rd	487.5	489.4	24.88	−1.91	−.077	.219	486.9	490.0	−3.12	−.126	−3.13	−.118
Spelling												
Pre-k	376.7	370.9	25.26	**5.81***	**.229**	**.000**	378.5	369.0	**9.53***	**.379**	**6.64***	**.255**
K	423.9	424.4	21.17	−0.56	−.027	.640	423.7	424.6	−0.92	−.043	−0.24	−.011
1st	459.4	461.6	20.95	**-2.23†**	**−.107**	**.073**	458.7	462.3	**−3.66†**	**−.175**	**−3.75†**	**−.173**
2nd	476.9	478.4	20.77	−1.58	−.076	.209	476.4	478.9	−2.59	−.125	−2.47	−.114
3rd	489.1	490.3	20.80	−1.19	−.057	.372	488.7	490.6	−1.95	−.094	−2.20	−.099
Oral comp												
Pre-k	452.0	451.3	16.90	0.69	.041	.355	452.2	451.1	1.13	.067	−0.48	−.032
K	465.9	465.3	15.43	0.58	.038	.446	466.1	465.1	0.96	.062	−0.40	−.029
1st	477.2	477.2	13.93	−0.03	−.002	.967	477.2	477.2	−0.05	−.004	−0.90	−.069
2nd	485.4	486.3	13.28	−0.89	−.067	.233	485.1	486.5	−1.46	−.110	−1.98	−.152
3rd	493.9	494.6	13.85	−0.66	−.048	.403	493.7	494.8	−1.09	−.079	**−2.48†**	**−.183**
Picture vocab												
Pre-k	462.8	459.3	18.27	**3.51***	**.192**	**.000**	463.9	458.2	**5.76***	**.316**	**3.31***	**.227**
K	472.3	470.8	11.89	**1.55***	**.130**	**.007**	472.8	470.3	**2.54***	**.214**	**2.16***	**.208**
1st	479.1	478.1	11.84	1.00	.084	.101	479.4	477.7	1.64	.139	**1.86†**	**.170**
2nd	484.9	484.6	11.50	0.27	.024	.653	485.0	484.5	0.44	.039	0.84	.076
3rd	491.2	490.9	11.58	0.34	.029	.606	491.3	490.8	0.55	.048	0.53	.047
Passage comp												
Pre-k	–	–	–	–	–	–	–	–	–	–	–	–
K	421.6	422.9	22.26	−1.24	−.056	.331	421.2	423.3	−2.03	−.091	−1.04	−.047
1st	457.6	458.1	19.59	−0.47	−.024	.684	457.4	458.2	−0.77	−.039	−0.27	−.013
2nd	474.2	474.5	16.64	−0.32	−.019	.742	474.1	474.6	−0.52	−.032	−0.60	−.036
3rd	483.8	483.9	16.34	−0.02	−.001	.984	483.8	483.9	−0.03	−.002	−0.22	−.013
Applied probs												
Pre-k	409.3	405.5	23.92	**3.81***	**.158**	**.001**	410.5	404.3	**6.25***	**.263**	1.57	.069
K	436.8	435.6	16.12	1.18	.073	.161	437.2	435.3	1.93	.120	1.63	.103
1st	458.0	456.9	16.08	1.15	.071	.190	458.4	456.5	1.89	.117	1.21	.076
2nd	473.9	474.6	17.04	−0.77	−.045	.435	473.6	474.9	−1.26	−.074	−1.49	−.085
3rd	485.2	486.8	18.20	−1.61	−.089	.139	484.7	487.3	−2.64	−.145	−1.53	−.084
Quant concepts												
Pre-k	422.1	418.9	16.04	**3.21***	**.199**	**.000**	423.1	417.8	**5.26***	**.330**	**3.99***	**.250**
K	448.7	448.3	13.54	0.32	.024	.660	448.8	448.2	0.53	.039	0.34	.025
1st	464.5	464.6	13.77	−0.08	−.006	.921	464.5	464.6	−0.13	−.009	−0.91	−.064
2nd	476.5	478.2	14.24	**−1.75***	**−.123**	**.039**	475.9	478.8	**−2.87***	**−.202**	**−3.43***	**−.234**
3rd	489.2	490.0	13.72	−0.82	−.060	.321	488.9	490.3	−1.35	−.098	−2.07	−.144
Calculation												
Pre-k	–	–	–	–	–	–	–	–	–	–	–	–
K	434.3	435.9	18.84	−1.53	−.081	.172	433.8	436.3	−2.51	−.133	−1.85	−.096
1st	462.2	462.7	15.75	−0.56	−.036	.580	462.0	462.9	−0.92	−.059	−0.37	−.022
2nd	476.8	476.9	12.17	−0.18	−.015	.817	476.7	477.0	−0.29	−.024	0.16	.012
3rd	490.4	491.0	13.92	−0.57	−.041	.528	490.2	491.2	−0.94	−.068	−0.76	−.052
	N = 697	*N* = 379					*N* = 780	*N* = 296				

*$p < .05$, †$p < 10$ for coefficients; significant coefficients and related estimates are also bolded.
Notes: Woodcock–Johnson W-scores. WJ Composite 6 is the mean of the W-scores across the content-specific scales except Passage Comprehension and Calculation. WJ Composite 8 is the mean of the W-scores including Passage Comprehension and Calculation, which were not administered at the end of the pre-k year.

[a] Covariate-adjusted means generated by the multilevel analysis models with covariates set at the grand means for the sample.

[b] Pooled treatment and control group standard deviations. There are minor variations between the pooled SDs for the ITT and TOT; the mean is presented here but effect sizes are computed on the exact values.

[c] Coefficients for treatment-control differences from OLS multilevel multiple imputation models with children nested in R-Lists and R-Lists nested in districts. Covariates are age at time of testing, male, Black, Hispanic, non-native English, mother's education, WJ Composite 6 pretest, Quantitative Concepts pretest, and pretest lag. The multiplier for ITT coefficients that estimates TOT coefficients is 1.6396 (see Table 4).

[d] Effect size: coefficient for the treatment-control difference divided by the pooled standard deviation.

[e] The 2SLS analysis model yields *p*-values for statistical significance that are the same for the ITT and TOT coefficients.

Please cite this article in press as: Lipsey, M. W., et al. Effects of the Tennessee Prekindergarten Program on children's achievement and behavior through third grade. *Early Childhood Research Quarterly* (2018), https://doi.org/10.1016/j.ecresq.2018.03.005

Table 8
Intent-to-treat (ITT) and treatment-on-treated (TOT) impact estimates for teacher ratings of preparedness for grade (ISS subsample).

Scale & year	ITT						TOT				TOT weighted	
	Treatment group mean[a]	Control group mean[a]	Pooled SD[b]	Coefficient[c] for T-C difference	Effect size[d]	*p*-value[e]	Treatment group mean[a]	Control group mean[a]	Coefficient[c] for T-C difference	Effect size[d]	Coefficient[c] for T-C difference	Effect size[d]
ACBR prep for grade												
K	4.53	4.38	1.46	**.147**[†]	**.100**	**.080**	4.57	4.33	**.242**[†]	**.166**	**.299**[*]	**.208**
1st	4.36	4.48	1.48	−.116	−.078	.205	4.33	4.52	−.190	−.128	−.238	−.160
2nd	4.31	4.41	1.45	−.095	−.066	.298	4.28	4.44	−.156	−.107	−.136	−.096
3rd	4.20	4.26	1.61	−.055	−.035	.574	4.19	4.28	−.091	−.056	−.082	−.050
	N = 697	*N* = 379					*N* = 780	*N* = 296				

[*] $p < .05$, [†] $p < .10$ for coefficients; significant coefficients and related estimates are also bolded.
Notes: ACBR Preparedness for Grade, scored as the mean of 3 items rated on 1–7 point scales.
[a] Covariate-adjusted means generated by the multilevel analysis models with covariates set at the grand means for the sample.
[b] Pooled treatment and control group standard deviations. There are minor variations between the pooled SDs for ITT and TOT; the mean is presented here but effect sizes are computed on the exact values.
[c] Coefficients for treatment-control differences from OLS multilevel multiple imputation models with children nested in R-Lists and R-Lists nested in districts. Covariates are age at time of rating, male, Black, Hispanic, non-native English, mother's education, WJ Composite6 pretest, and Quantitative Concepts pretest. The multiplier for the ITT coefficients that estimate the TOT coefficients is 1.6396 (see Table 4).
[d] Effect size: coefficient for the treatment-control difference divided by the pooled standard deviation.
[e] The 2SLS analysis model yields *p*-values for statistical significance that are the same for the ITT and TOT coefficients.

Table 9
Intent-to-treat (ITT) and treatment-on-treated (TOT) impact estimates for the state achievement tests (RCT sample).

Analysis & subject	ITT						TOT			
	Treatment group mean[a]	Control group mean[a]	Pooled SD[b]	Coefficient[c] for T-C difference	Effect size[d]	*p*-value[e]	Treatment group mean[a]	Control group mean[a]	Coefficient[c] for T-C difference	Effect size[d]
Multiple imputation										
Reading	744.9	747.3	35.24	−2.34	−.066	.130	743.9	748.3	−4.45	−.126
Mathematics	755.9	760.4	36.59	**−4.46**[*]	**−.122**	**.006**	753.9	762.4	**−8.48**[*]	**−.232**
Science	749.0	752.8	36.34	**−3.87**[*]	**−.106**	**.016**	747.2	754.6	**−7.34**[*]	**−.202**
	N = 1852	*N* = 1138		*N* = 2990			*N* = 1997	*N* = 993	*N* = 2990	
Observed values										
Reading	746.2	748.1	34.33	−1.82	−.053	.214	745.4	748.9	−3.45	−.100
Mathematics	757.1	761.0	31.55	**−3.86**[*]	**−.140**	**.011**	755.4	762.7	**−7.34**	**−.206**
Science	750.1	753.2	35.32	**−3.17**[*]	**−.090**	**.033**	748.7	754.7	**−6.02**	**−.170**
	N = 1505	*N* = 935		*N* = 2440			*N* = 1638	*N* = 802	*N* = 2440	
Weighted imputation										
Reading	746.4	748.6	34.96	−2.24	−.064	.356	745.3	749.6	−4.26	−.122
Mathematics	756.3	759.7	36.59	−3.35	−.092	.162	754.8	761.2	−6.36	−.174
Science	750.7	753.9	36.60	−3.13	−.085	.189	749.3	755.3	−5.94	−.162
	N = 1852	*N* = 1138		*N* = 2990			*N* = 1997	*N* = 993	*N* = 2990	

[*] $p < .05$, [†] $p < .10$ for coefficients; significant coefficients and related estimates are also bolded.
Notes: TCAP state achievement scaled scores.
[a] Covariate-adjusted means generated by the multilevel analysis models with covariates set at the grand means for the sample.
[b] Pooled treatment and control group standard deviations. There are minor variations between the pooled SDs for ITT and TOT; the mean is presented here but effect sizes are computed on the exact values.
[c] Coefficients for treatment-control differences from OLS multilevel models with children nested in R-Lists and R-Lists nested in districts. Covariates are age, male, Black, Hispanic, and non-native English. The multipliers for the ITT coefficients that estimate the TOT coefficients are 1.8993 with multiple imputation and 1.8990 with observed values.
[d] Effect size: coefficient for the treatment-control difference divided by the pooled standard deviation.
[e] The 2SLS analysis model yields *p*-values for statistical significance that are the same for the ITT and TOT coefficients.

VPK treatment children in all three subject areas. Those differences were statistically significant for math and science in the analyses with multiple imputation and observed scores, but not for reading. Negative effect estimates also appeared in the weighted analyses that generalized to the statewide VPK population, but they were smaller and fell short of statistical significance so are best characterized as null effects.

It is noteworthy that the pattern of null and negative VPK effects on the state achievement tests for the RCT sample is substantially similar to that found on the WJ achievement measures for the ISS subsample (Table 7) and consistent with teachers' ratings of preparedness for grade level work (Table 8). Across all these achievement-related outcomes, there were no statistically significant differences past the kindergarten year that favored the VPK participants. Moreover, the direction of the differences in those later grades was overwhelmingly negative, indicating that the VPK participants did not perform as well as the control children, with a number of those differences reaching statistical significance.

5.1.5. *Achievement effects for different subgroups of children*

Further analysis of the various VPK achievement effects was conducted to identify any differential effects for demographic subgroups of children at the end of the pre-k year and later.

5.1.5.1. Differential effects at the end of the pre-k year. Achievement data at the end of the pre-k year were available on the WJ measures and teacher ratings, but only for the ISS subsample. Analyses like those reported in Table 7 for main effects on the WJ achievement measures were repeated for the WJ Composite 6 summary measure with addition of terms for the interactions of the ITT treatment condition with the baseline variables for

Please cite this article in press as: Lipsey, M. W., et al. Effects of the Tennessee Prekindergarten Program on children's achievement and behavior through third grade. *Early Childhood Research Quarterly* (2018), https://doi.org/10.1016/j.ecresq.2018.03.005

Table 10
ITT and TOT subgroup impact estimates for the WJ Composite 6 measure at the end of the pre-k year (ISS subsample).

Moderator variable and subgroup	ITT						TOT			
	Treatment group mean[a] (N)	Control group mean[a] (N)	Pooled SD[b]	Treatment-control difference[c]	Effect size[d]	*p*-value[e]	Treatment group mean[a] (N)	Control group mean[a] (N)	Treatment-control difference[c]	Effect size[d]
Native English			16.80			**.001***				
Yes	409.9 (592)	406.5 (269)		3.35	.197		410.9 (654)	405.5 (207)	5.49	.330
No	412.7 (105)	398.9 (110)		13.79	.813		417.1 (126)	394.5 (89)	22.61	1.36
WJ Composite 6 pretest			16.80			**.001***				
Lower scores	406.0 (232)	396.2 (127)		9.86	.581		409.2 (249)	393.0 (111)	16.17	.972
Higher scores	416.5 (465)	409.2 (252)		7.27	.429		418.8 (531)	406.9 (185)	11.93	.717
English × Comp 6 pretest			16.80			**.056†**				
Yes/low score	405.6 (151)	400.9 (50)		4.75	.280		407.1 (158)	399.3 (43)	7.78	.468
Yes/high score	414.1 (441)	412.2 (219)		1.95	.115		414.8 (496)	411.6 (164)	3.19	.192
No/low score	406.5 (81)	391.5 (77)		14.98	.883		411.3 (91)	386.7 (68)	24.56	1.48
No/high score	418.9 (24)	406.3 (33)		12.60	.743		422.9 (35)	402.3 (21)	20.66	1.24

*$p < .05$, †$p < .10$ for coefficients; significant coefficients and related estimates are also bolded.
Notes: Woodcock–Johnson Composite 6 W-scores. To illustrate the magnitude of the relations, pretest WJ Composite 6 scores are grouped into the lowest tertile vs. the two higher tertiles.
[a] Covariate-adjusted means generated by the multilevel analysis models with covariates set at the grand means for the sample; OLS multilevel multiple imputation models with children nested in R-Lists and R-Lists nested in districts. Covariates are age at time of testing, male, Black, Hispanic, non-native English, mother's education, WJ Composite 6 pretest, Quantitative Concepts pretest, and pretest lag.
[b] Pooled treatment and control group standard deviations for the aggregate treatment and control groups to facilitate comparison across subgroups and with the main effects in Table 7. There are minor variations between the pooled SDs for the ITT and TOT; the mean is presented here but effect sizes are computed on the exact values.
[c] Difference between the treatment and control covariate-adjusted means. The multiplier for the ITT differences that estimate the TOT differences is 1.6396 (see Table 4).
[d] Effect size: treatment-control difference divided by the pooled standard deviation.
[e] *p*-values for the respective interaction terms in the analysis model. The 2SLS analysis model yields *p*-values for statistical significance that are the same for the ITT and TOT coefficients.

children's age, gender, race, whether English was the native language, mother's education, and performance on the WJ Composite 6 pretest. Larger positive and statistically significant VPK effects were found for non-native English speakers compared with native English speakers, and for children who scored lowest on the Composite 6 pretest achievement measure relative to those with higher scores. Table 10 summarizes these results, showing the treatment and control group means for each subgroup and the associated effect sizes, with subgroups on the Composite 6 measure created by contrasting children who scored in the lower tertile of the distribution with those with higher scores.

Table 10 also reports the breakout for the subgroups of the three-way interaction that combines VPK treatment status with both the native English and Composite 6 pretest variables. The largest VPK effects were for non-native English speakers with low pretest achievement scores. Conversely, the smallest effects were for native English speaking children with higher pretest achievement scores. Note that the Composite 6 pretest is correlated with other moderator variables that did not emerge so strongly in their own right in these analyses. In particular, Composite 6 scores are significantly lower for children who are younger, have mothers with less education, are Hispanic (the largest concentration of non-native English speakers), or are male.

Differential effects on teachers' ratings of preparedness for grade-level work at the beginning of kindergarten were investigated with the same procedure. None of the interactions of the baseline variables with ITT treatment condition was statistically significant for those ratings.

5.1.5.2. Differential effects after the end of the pre-k year. The large gain on the WJ Composite 6 achievement measure during the pre-k year by VPK children who were not native English speakers and/or had low pretest scores on that measure (Table 10) might propel sustained achievement advantages for those subgroups in the later grades. However, analysis of the interactions of ITT treatment condition with age, gender, race, native English, mothers' education, and the WJ Composite 6 pretest revealed no statistically significant differential effects on the WJ Composite 6 achievement measure at the end of kindergarten or later through the 2nd grade year. A small statistical interaction with the WJ Composite 6 pretest ($p < .07$) emerged in the 3rd grade year, but in the opposite direction from that in the pre-k year – the initially lowest scoring children whose participation in VPK produced especially large gains on the Composite 6 achievement measure during pre-k had fallen behind the control children on that measure by the 3rd grade year. The magnitude of the effect can again be illustrated by contrasting the VPK effect on the Composite 6 outcome for children scoring in the lowest tertile with those in the upper tertiles on that measure at pretest. For the ITT comparison, that representation yields effect sizes in the 3rd grade year of −.06 for the initially lowest scoring group and −.01 for the higher scoring group. For the TOT comparison, the corresponding effect sizes were −.11 and −.01.

Teacher ratings of preparedness for grade level work in the years after pre-k also showed negative differential effects for the children with lower pretest scores on WJ Composite 6 achievement. Those effects were statistically significant for teacher ratings at the end of the 1st, 2nd, and 3rd grade years. The effect sizes on those ratings for children in the lowest tertile of the Composite 6 pretest compared to those with higher scores for the ITT comparison (TOT effect sizes in parentheses) were −.27 vs .05 (−.44 vs .09) at the end of the 1st grade year, −.26 vs .06 (−.43 vs .11) at the end of 2nd grade, and −.13 vs .04 (−.21 vs .06) at the end of 3rd grade.

Examined individually, there were a number of other statistically significant interactions that indicated differential effect on the ratings of preparedness for grade after the pre-k year, in particular, for Black and Hispanic children, non-native English speakers, and children with less educated mothers. However, all those variables are correlated with the baseline Composite 6 achievement pretest. When the differential effects associated with that pretest were accounted for in the analyses, none of these other interaction terms remained statistically significant.

Further exploration was made of differential effects on the state achievement tests in 3rd grade with the full RCT sample. The data for that sample do not include pretest achievement measures, so no direct comparison with the results from the ISS subsample can be made. The interactions of ITT treatment condition with age, gender, race, and native English were tested, but none was statistically significant for reading, math, or science.

Please cite this article in press as: Lipsey, M. W., et al. Effects of the Tennessee Prekindergarten Program on children's achievement and behavior through third grade. Early Childhood Research Quarterly (2018), https://doi.org/10.1016/j.ecresq.2018.03.005

Table 11
ITT and TOT impact estimates for retention in grade (RCT sample).

Analysis & grade year	ITT						TOT			
	Treatment group mean[a]	Control group mean[a]	Pooled SD[b]	Coefficient[c] for T-C difference	Effect size[d]	*p*-value[e]	Treatment group mean[a]	Control group mean[a]	Coefficient[c] for T-C difference	Effect size[d]
Multiple imputation										
K	.052	.066	.227	−.014	−.064	.108	.045	.072	−.027	−.121
K-1st	.109	.103	.299	.006	.019	.622	.111	.100	.011	.037
K-2nd	.126	.128	.321	−.001	−.004	.921	.126	.128	−.002	−.008
K-3rd	.137	.137	.335	.000	.001	.976	.137	.137	.001	.002
	N = 1852	*N* = 1138					*N* = 1997	*N* = 993		
Observed values										
K	.050	.065	.224	**−.015**†	**−.067**	**.084**	.043	.072	**−.029**†	**−.128**
K-1st	.107	.099	.296	.008	.025	.516	.111	.096	.014	.048
K-2nd	.123	.121	.316	.002	.006	.870	.126	.120	.004	.012
K-3rd	.130	.126	.324	.003	.010	.801	.137	.125	.006	.019
	N = 1764–1807	*N* = 1067–1113					*N* = 1902–1953	*N* = 929–967		
Weighted imputation										
K	.055	.064	.222	−.009	−.043	.309	.050	.068	−.018	−.081
K-1st	.117	.103	.302	.015	.048	.247	.124	.096	.028	.091
K-2nd	.141	.120	.330	.021	.064	.361	.150	.111	.040	.121
K-3rd	.154	.125	.344	.028	.083	.221	.166	.112	.054	.157
	N = 1852	*N* = 1138					*N* = 1997	*N* = 993		

*$p < .05$, †$p < .10$ for coefficients; significant coefficients and related estimates are also bolded.
Notes: Cumulative values across the grades indicated; e.g., K-2nd refers to retention in any grade from K through 2nd.
[a] Covariate-adjusted means generated by the multilevel analysis models with covariates set at the grand means for the sample.
[b] Pooled treatment and control group standard deviations. There are minor variations between the pooled SDs for the ITT and TOT; the mean is presented here but effect sizes are computed on the exact values.
[c] Coefficients for the treatment-control differences from OLS multilevel models with children nested in R-Lists and R-Lists nested in districts. Covariates are age, male, Black, Hispanic, and non-native English. The multiplier for the ITT coefficient that estimates the TOT coefficient is 1.8993 with multiple imputation and varies from 1.9004 to 1.9051 with observed values.
[d] Effect size: coefficient for the treatment-control difference divided by the pooled standard deviation.
[e] The 2SLS analysis model yields *p*-values for statistical significance that are the same for the ITT and TOT coefficients.

5.1.6. Retention in grade (RCT)

Because retained students continue to lag behind their peers in subsequent grades, we examined cumulative retention – the proportion of students in each successive school year below the expected grade level for that year irrespective of the year in which they had been retained. Table 11 reports the retention rates for students in the treatment and control conditions of the RCT sample. For example, a K-3rd retention rate of .13 means that 13% of the children were retained in the 3rd grade or a prior year.

The top panel of Table 11 reports results from multilevel OLS analysis with multiple imputation of missing data. Substantially similar retention rate estimates were found with analyses of only observed values (second panel of Table 11), and with multilevel logistic regression rather than OLS for the binary retention outcomes each year (not shown). The lower panel of Table 11 presents the results of the retention analyses with application of the weighting function that yields estimates of the expected results if those in the upper panel are extended to the statewide population of VPK programs and participating children.

As Table 11 indicates, a somewhat lower proportion of VPK children was retained in kindergarten than the corresponding control children, but those differences were statistically significant only at the $p < .10$ level and only for the observed data. Retention through 1st grade reversed the proportions, although not statistically significant. That is, the retention rates for the VPK groups caught up with those for the control groups by 1st grade and there were no significant differences between the groups thereafter.

5.1.6.1. Differential effects on retention. Further analyses explored differential VPK effects on retention for subgroups of children identified by the demographic variables for the RCT sample (age, gender, race, and native English). The only statistically significant interaction with ITT treatment condition ($p = .026$) was for age in kindergarten. While younger children were more likely to be retained throughout the K-3rd grade period, the younger VPK participants were less likely to be retained in kindergarten than their counterparts in the control group. The retention rate in kindergarten for the VPK children below the mean age of their classmates was 7.0% versus 11.2% for the corresponding control children in the TOT comparison; the analogous difference for children above the mean age was 2.2% vs. 3.2%.

5.2. Classroom and school behavior

5.2.1. Teacher ratings (ISS)

For the ISS subsample, teachers rated children's behavior and performance in the classroom on the Cooper–Farran Work-Related Skills and Interpersonal Skills scales, and the ACBR Peer Relations, Behavior Problems, and Feelings About School scales. Table 12 shows the effect estimates on these measures through the 3rd grade year. There is no consistent direction in the effect sizes for these ratings and none are statistically significant with one exception. The 1st grade teachers' ratings of children's feelings about school was statistically significant with VPK children having less positive attitudes toward school than control children. Application of the weighting function that extrapolated the results in Table 12 to the statewide VPK population (not shown) produced comparable findings with negative VPK effects ($p < .10$) on teacher ratings of feelings about school in the 1st and 2nd grade years.

5.2.2. Disciplinary actions (RCT)

School records in the state database report expulsions and in-school and out-of-school suspensions. The frequency of such events is low, especially in any one school year. The outcome variable used here thus aggregates across the K-3rd grade years to indicate whether there were any recorded actions (yes/no) during that period. The overall rates are further subdivided into those for less serious infractions such as breaking school rules and related admin-

Please cite this article in press as: Lipsey, M. W., et al. Effects of the Tennessee Prekindergarten Program on children's achievement and behavior through third grade. *Early Childhood Research Quarterly* (2018), https://doi.org/10.1016/j.ecresq.2018.03.005

Table 12
ITT and TOT impact estimates for teacher ratings of classroom behavior (ISS subsample).

	ITT						TOT	
Rating scale & year	Treatment group mean[a]	Control group mean[a]	Pooled SD[b]	Coefficient[c] for T-C difference	Effect size[d]	*p*-value[e]	Coefficient[c] for T-C difference	Effect size[d]
CF Interpersonal Skills								
K (fall)	5.78	5.69	.96	.095	.099	.150	.156	.162
1st (spring)	5.80	5.88	1.01	−.075	−.074	.272	−.123	−.122
2nd (spring)	5.87	5.89	.97	−.020	−.021	.763	−.033	−.034
3rd (spring)	6.00	5.94	1.21	.060	.049	.385	.098	.081
CF Work-Related Skills								
K (fall)	4.99	4.93	1.16	.051	.044	.475	.083	.072
1st (spring)	5.03	5.17	1.25	−.133	−.107	.108	−.219	−.175
2nd (spring)	5.09	5.15	1.19	−.064	−.054	.408	−.106	−.089
3rd (spring)	5.23	5.20	1.67	.031	.019	.721	.051	.031
ACBR Peer Relations								
K (fall)	5.21	5.17	1.01	.046	.046	.492	.076	.075
1st (spring)	5.26	5.24	1.11	.017	.016	.827	.028	.026
2nd (spring)	5.24	5.20	1.11	.033	.030	.662	.054	.049
3rd (spring)	5.16	5.22	1.35	−.061	−.045	.460	−.101	−.074
ACBR Behavior Problems								
K (fall)	.42	.46	.50	−.037	−.076	.257	−.061	−.124
1st (spring)	.43	.40	.49	.031	.062	.340	.050	.102
2nd (spring)	.40	.42	.49	−.023	−.047	.485	−.038	−.077
3rd (spring)	.41	.39	.49	.013	.027	.682	.022	.045
ACBR Feelings about School								
K (fall)	1.62	1.60	.49	.018	.037	.584	.029	.061
1st (spring)	1.53	1.60	.50	**−.072**[*]	**−.146**	**.032**	**−.119**[*]	**−.239**
2nd (spring)	1.51	1.55	.50	−.039	−.077	.266	−.063	−.127
3rd (spring)	1.47	1.47	.50	−.002	−.004	.952	−.003	−.007
	N = 697	*N* = 379					*N* = 780 & 296	

[*] $p < .05$, [†] $p < .10$ for coefficients; significant coefficients and related estimates are also bolded.
Notes: CF Interpersonal Skills, CF Work-Related Skills, and ACBR Peer Relations scored as means of 1-7 point scales. ACBR Behavior Problems scored as 1 for any problem indication, 0 otherwise. ACBR Feelings about School is scored 2 for ratings near the highest possible, 1 for lower ratings.
[a] Covariate-adjusted means generated by the multilevel analysis models with covariates set at the grand means for the sample.
[b] Pooled treatment and control group standard deviations. There were minor variations between the pooled SDs for the ITT and TOT; the mean is presented here but effect sizes were computed on the exact values.
[c] Coefficients for the treatment-control differences from OLS multilevel multiple imputation models with children nested in R-Lists and R-Lists nested in districts. Covariates are age at time of rating, male, Black, Hispanic, non-native English, mother's education, WJ Composite6 pretest, and Quantitative Concepts. The multiplier for ITT coefficients that estimates the TOT coefficients is 1.6396 (see Table 4).
[d] Effect size: coefficient for the treatment-control difference divided by the pooled standard deviation.
[e] The 2SLS analysis model yields *p*-values for statistical significance that are the same for the ITT and TOT coefficients.

istrative matters and more serious infractions such as fighting, bullying, and bringing a weapon to school. Table 13 summarizes the results for the primary analysis with multiple imputation for missing values, analysis with only observed values, and the weighted analysis that generalizes to the statewide population of programs and children. All those analyses found a higher frequency of violations of school rules for VPK participants that was marginally significant ($p < .10$). The weighted analysis also showed a significantly greater frequency of major infractions and all infractions taken together for the VPK participants.

5.2.3. Attendance (RCT)

Attendance rates for children in the RCT sample were derived from state data for kindergarten through 3rd grade as the proportion of instructional days for which a child did not have a recorded absence. As shown in Table 14, those rates were universally high, around 95%, and no significant differences were found between the VPK treatment and control groups for any year. The results using only observed data and applying the weighting function to estimate statewide effects are not shown but are substantially similar.

5.2.4. Classroom and school behavior effects for different subgroups of children

Moderator analyses were conducted for the classroom and school outcomes to identify any differential effects for demographic subgroups of children. To accomplish this, the analyses of main effects reported above were repeated with the addition of interaction terms for the ITT treatment condition by subgroup relationships at each wave of data collection.

5.2.5. Differential effects on teacher ratings (ISS)

Two patterns of differential VPK effects emerged for teacher ratings when statistical interaction effects were tested for age, gender, race/ethnicity, mother's education, and pretest achievement level in the ISS subsample. In 2nd grade, significant interaction effects showed that male VPK participants received more positive teacher ratings than males in the control group in several domains. The rating scales that showed these differential effects (ITT effect sizes for treatment-control differences for males vs. those for females in parentheses, followed by the corresponding TOT effect sizes) are as follows: Cooper–Farran Interpersonal Skills (.11 vs −.14; .18 vs −.23), ACBR Peer Relations (.16 vs −.08; .25 vs −.13), ACBR Behavior Problems (-.19 vs .08; −.32 vs .14), and ACBR Feelings About School (.09 vs −.23; .14 vs −.37). Conversely, female VPK participants received less positive ratings than females in the control group on these scales, as indicated by the second effect size in each pair. In 3rd grade, a similar differential favoring male students appeared for ACBR Peer Relations (.09 vs −.16; .14 vs −.26) and Cooper–Farran Work-Related Skills (.11 vs −.06; .18 vs −.10).

Teacher ratings also showed differential VPK effects for Black children in contrast to their mainly White and Hispanic counterparts. In 1st and 3rd grade statistically significant interactions revealed higher ratings for Black VPK participants than for Black control children on Cooper–Farran Work-Related Skills (1st grade: .11 vs −.17; .17 vs −.28. 3rd grade: .20 vs −.04; .33 vs −.06). Con-

Please cite this article in press as: Lipsey, M. W., et al. Effects of the Tennessee Prekindergarten Program on children's achievement and behavior through third grade. *Early Childhood Research Quarterly* (2018), https://doi.org/10.1016/j.ecresq.2018.03.005

Table 13
ITT and TOT impact estimates for disciplinary actions (RCT sample).

Analysis & outcome	ITT						TOT			
	Treatment group mean[a]	Control group mean[a]	Pooled SD[b]	Coefficient[c] for T-C difference	Effect size[d]	*p*-value[e]	Treatment group mean[a]	Control group mean[a]	Coefficient[c] for T-C difference	Effect size[d]
Multiple imputation										
School rules	.064	.049	.225	**.015**†	**.065**	**.098**	.071	.043	**.028**†	**.123**
Major offenses	.034	.034	.183	−.001	−.003	.940	.034	.035	−.001	−.006
All offenses	.080	.071	.261	.009	.033	.403	.084	.067	.016	.062
	N = 1852	*N* = 1138					*N* = 1997	*N* = 993		
Observed values										
School rules	.068	.051	.231	**.017**†	**.072**	**.072**	.075	.044	**.031**†	**.136**
Major offenses	.036	.036	.189	.000	.002	.961	.036	.036	.001	.004
All offenses	.085	.075	.269	.010	.037	.359	.089	.070	.019	.069
	N = 1705	*N* = 1034					*N* = 1849	*N* = 890		
Weighted imputation										
School rules	.070	.050	.244	**.020**†	**.080**	**.053**	.079	.042	**.037**†	**.152**
Major offenses	.050	.021	.207	**.029***	**.138**	**.001**	.062	.008	**.054***	**.263**
All offenses	.101	.059	.290	**.042***	**.144**	**.000**	.119	.040	**.079***	**.273**
	N = 1852	*N* = 1138					*N* = 1997	*N* = 993		

*$p < .05$, †$p < .10$ for coefficients; significant coefficients and related estimates are also bolded.
Notes: School rules: violations of school rules or other administrative issues; major offenses: fighting, bullying, weapon in school, and the like; all offenses: total across school rule and major offenses categories. These are coded for whether there is any infraction recorded in school records (1 = yes, 0 = no) cumulatively from K through the 3rd grade year.
[a] Covariate-adjusted means generated by the multilevel analysis models with covariates set at the grand means for the sample.
[b] Pooled treatment and control group standard deviations. There were minor variations between the pooled SDs for the ITT and TOT; the mean is presented here but effect sizes were computed on the exact values.
[c] Coefficients for the treatment-control differences from OLS multilevel models with children nested in R-Lists and R-Lists nested in districts. Covariates are age, male, Black, Hispanic, and non-native English. The multiplier for ITT coefficients that estimates TOT coefficients is 1.8993 with multiple imputation and 1.8765 with observed values.
[d] Effect size: coefficient for the treatment-control difference divided by the pooled standard deviation.
[e] The 2SLS analysis model yields *p*-values for statistical significance that are the same for the ITT and TOT coefficients.

versely, for White and Hispanic children, VPK participants were rated lower on this scale than White and Hispanic control children. A similar relationship favoring Black VPK participants relative to their counterparts in the control group was found in 3rd grade on the Cooper–Farran Interpersonal Skills scale (.24 vs −.01; .39 vs −.02). In 2nd grade, another statistically significant differential indicated that Black VPK participants received higher ratings than Black control group children on the ACBR Feelings About School scale while the contrasting group of White and Hispanic participants received lower ratings than their corresponding controls (.17 vs −.15; .27 vs −.25). There were only a few other statistically significant interaction effects of the many tested, none of which involved as many of the teacher ratings for particular subgroups as those for male and Black children.

5.2.6. *Differential effects on disciplinary actions and attendance (RCT)*

No statistically significant differential effects for demographic subgroups were found for disciplinary outcomes. For attendance, however, VPK effects were more positive and statistically significant for male and for Black children in the 2nd grade, and for Black children in the 3rd grade. Conversely, female and non-Black VPK participants had lower attendance rates than the corresponding control children. These gender and race/ethnic interactions echo some of those described above for teacher ratings for male and Black children. Whereas, there was no evidence that VPK effects on achievement were stronger for those subgroups, some of their school and classroom behavior past the pre-k year does appear to have been positively affected by their VPK participation.

Table 14
ITT and TOT impact estimates for attendance (RCT sample).

Analysis & grade year	ITT						TOT			
	Treatment group mean[a]	Control group mean[a]	Pooled SD[b]	Coefficient[c] for T-C difference	Effect size[d]	*p*-value[e]	Treatment group mean[a]	Control group mean[a]	Coefficient[c] for T-C difference	Effect size[d]
Multiple imputation										
K	.947	.950	.042	−.002	−.055	.158	.946	.951	−.004	−.104
1st	.954	.955	.038	.000	−.008	.838	.954	.955	−.001	−.015
2nd	.959	.960	.036	−.002	−.042	.286	.958	.961	−.003	−.081
3rd	.962	.963	.042	−.001	−.032	.426	.961	.964	−.003	−.062
	N = 1852	*N* = 1138					*N* = 1997	*N* = 993		

*$p < .05$, †$p < .10$.for coefficients; significant coefficients and related estimates are also bolded.
[a] Covariate-adjusted means generated by the multilevel analysis models with covariates set at the grand means for the sample.
[b] Pooled treatment and control group standard deviations. There are minor variations between the pooled SDs for the ITT and TOT; the mean is presented here but effect sizes are computed on the exact values.
[c] Coefficients for the treatment-control differences from OLS multilevel multiple imputation models with children nested in R-Lists and R-Lists nested in districts. Covariates are age, male, Black, Hispanic, and non-native English. The multiplier for ITT coefficients that estimates TOT coefficients is 1.8993 (see Table 4).
[d] Effect size: coefficient for the treatment-control difference divided by the pooled standard deviation.
[e] The 2SLS analysis model yields *p*-values for statistical significance that are the same for the ITT and TOT coefficients.

Please cite this article in press as: Lipsey, M. W., et al. Effects of the Tennessee Prekindergarten Program on children's achievement and behavior through third grade. Early Childhood Research Quarterly (2018), https://doi.org/10.1016/j.ecresq.2018.03.005

Table 15
ITT and TOT impact estimates for special education IEPs (RCT sample).

Outcome, analysis, & grade year	ITT						TOT			
	Treatment group mean[a]	Control group mean[a]	Pooled SD[b]	Coefficient[c] for T-C difference	Effect size[d]	*p*-value[e]	Treatment group mean[a]	Control group mean[a]	Coefficient[c] for T-C difference	Effect size[d]
Any IEP except gifted or physical disability (multiple imputation)										
K	.129	.096	.304	**.034***	**.111**	**.004**	.144	.080	**.064***	**.211**
1st	.138	.106	.321	**.031***	**.097**	**.013**	.152	.092	**.059***	**.185**
2nd	.135	.113	.331	**.022**†	**.068**	**.088**	.145	.103	**.043**†	**.128**
3rd	.133	.106	.331	**.027***	**.082**	**.039**	.146	.094	**.052***	**.156**
	N = 1852	*N* = 1138					*N* = 1997	*N* = 993		
Any IEP except gifted or physical disability (observed values)										
K	.130	.096	.305	**.034***	**.111**	**.004**	.145	.081	**.065***	**.212**
1st	.138	.107	.322	**.031***	**.097**	**.013**	.153	.093	**.060***	**.186**
2nd	.139	.116	.330	**.023**†	**.070**	**.078**	.150	.106	**.044**†	**.134**
3rd	.137	.111	.331	**.026**†	**.079**	**.051**	.149	.099	**.049**†	**.150**
Intellectually gifted (observed values)										
K	.002	.004	.049	−.002	−.047	.232	.001	.005	−.004	−.090
1st	.002	.009	.072	**−.007***	**−.091**	**.022**	−.001	.012	**−.012***	**−.173**
2nd	.005	.010	.092	−.005	−.054	.177	.003	.012	−.009	−.103
3rd	.008	.014	.109	−.006	−.054	.180	.005	.017	−.011	−.103
Speech/language impairment (observed values)										
K	.123	.087	.295	**.036***	**.122**	**.001**	.139	.071	**.068***	**.231**
1st	.121	.094	.304	**.026***	**.086**	**.026**	.132	.082	**.050***	**.165**
2nd	.115	.101	.298	.014	.048	.230	.121	.094	.027	.091
3rd	.098	.092	.281	.007	.024	.554	.101	.089	.013	.045
Specific learning & intellectual disabilities (observed values)										
K	.015	.014	.116	.001	.000	.754	.016	.013	.003	.000
1st	.029	.017	.148	**.013***	**.084**	**.033**	.035	.011	**.024***	**.160**
2nd	.048	.034	.196	**.014**†	**.073**	**.068**	.054	.027	**.027**†	**.139**
3rd	.068	.042	.228	**.026***	**.113**	**.006**	.079	.030	**.049***	**.214**
	N = 1745–1825	*N* = 1056–1121					*N* = 1881–1980	*N* = 920–965		

* $p < .05$, † $p < .10$ for coefficients; significant coefficients and related estimates are also bolded.
Notes: IEP = Individualized Educational Program as the formal special education designation.
[a] Covariate-adjusted means generated by the multilevel analysis models with covariates set at the grand means for the sample.
[b] Pooled treatment and control group standard deviations. There are minor variations between the pooled SDs for the ITT and TOT; the mean is presented here but effect sizes are computed on the exact values.
[c] Coefficients for the treatment-control differences from OLS multilevel models with children nested in R-Lists and R-Lists nested in districts. Covariates are age, male, Black, Hispanic, and non-native English. The multiplier for ITT coefficients that estimates TOT coefficients is 1.8993 with multiple imputation and varies from 1.8950 to 1.9055 with observed values.
[d] Effect size: coefficient for the treatment-control difference divided by the pooled standard deviation.
[e] The 2SLS analysis model yields *p*-values for the statistical significance that are the same for the ITT and TOT coefficients.

5.3. Special education (RCT)

Special education services defined by an Individualized Education Program (IEP) are often of interest as pre-k outcomes. However, for pre-k programs like VPK that are administered by the department of education, located mainly in schools, and staffed by licensed teachers, it is not entirely clear how to interpret special education placement rates. In those circumstances, pre-k students are screened for special needs and special education services may be provided as part of the pre-k program, an opportunity not generally available to the control children who are not in VPK classrooms. Once begun, special education services typically continue for some years, ending only when a formal determination is made that they are no longer needed.

For VPK participants in the RCT sample, 9.4% had at least one IEP recorded in the state education database during the pre-k year in a category other than physical disabilities (0.2% of the children) and intellectually gifted (no children). Of these children, 93% had an IEP for speech or language impairment and 10% had an IEP for an intellectual disability (some had more than one IEP). For nearly all of the children given an IEP in pre-k, it was continued into the kindergarten year (93%) and, in most cases, for some time beyond kindergarten (80%).

This early identification of VPK participants for special education services, and continuation of that status into later grades, creates an asymmetry that confounds any attempt to interpret special education differences between VPK treatment and control groups in later grades. Such differences might represent the impact of pre-k, including early provision of special education services, on later need for such services. But, they may also represent carryover of the special education designations from the pre-k year for VPK participants with equally needy control children not receiving such designations until kindergarten at the earliest. These different effects cannot be disentangled in the data available for this study, but given the rate of special education designations in pre-k and the high proportion carried forward into the later grades, there is little basis to doubt that carryover plays a role.

With this ambiguity about how to interpret the results in mind, Table 15 reports the proportions of VPK treatment and control children with recorded IEPs from K to the 3rd grade year. The first two panels report IEPs other than those relating to physical disabilities or giftedness for the primary analysis with multiple imputation and the parallel analysis with only observed values. The remaining panels break out the different special education categories for analysis with observed values, omitting physical disabilities.

Table 15 shows that a higher proportion of VPK children had IEPs than control children in every year with differences that are often statistically significant. The one exception is the intellectually gifted category, where the proportions are low but favor the control groups. The results for speech or language impairments are especially noteworthy. This is the overwhelmingly dominant category for the early IEP designations for VPK participants and continues to be so through the 3rd grade year. The treatment-control differences shrink somewhat over those years as a result of small declining

Please cite this article in press as: Lipsey, M. W., et al. Effects of the Tennessee Prekindergarten Program on children's achievement and behavior through third grade. *Early Childhood Research Quarterly* (2018), https://doi.org/10.1016/j.ecresq.2018.03.005

proportions in the VPK treatment groups and relatively constant proportions in the control groups. If continued into later years that trend could eventually yield a more favorable outcome for VPK. In contrast, the category of specific learning and intellectual disabilities, albeit small, shows an increasing trend for both groups that increases more sharply for VPK participants.

Differential VPK effects on special education status were explored for the demographic subgroups in the RCT sample. Interactions between the ITT treatment condition and age, gender, Black, Hispanic, and whether a non-native English speaker were tested for the overall outcome of any IEP other than gifted or physical disabilities at each K-3rd grade year. A weak pattern of mostly $p < .10$ interaction effects emerged for males and Black children. In the early grades (K and 1st), VPK participation was associated with more IEPs for males relative to males in the control group; conversely female participants had fewer IEPs relative to females in the control group. But those difference diminished in 2nd and 3rd grade. For Black children, VPK participants had fewer IEPs from 1st through 3rd grade than their counterparts in the control group.

6. Discussion

The results of this study of the Tennessee Voluntary Pre-K program have addressed our initial research questions while raising still others. The first of those research questions is about the effects of VPK participation on literacy, language, and math skills, and classroom behavior, by kindergarten entry. The second question is whether any advantage of VPK participation carries forward to enhance performance in later grades.

For the ISS subsample we found that children who participated in VPK experienced considerably greater gains in literacy, language, and math skills during the pre-k year than the control children, and that this difference was recognized as greater preparedness for grade level work by kindergarten teachers at the beginning of the following year.

However, those positive VPK effects on achievement largely disappeared by the end of kindergarten with children in the control group catching up to the VPK participants. Moreover, by second grade the performance of the control children surpassed that of the VPK participants on some achievement measures. This pattern was echoed on the 3rd grade state achievement tests for the full RCT sample. VPK participants scored lower on the reading, math, and science tests than the control children with differences that were statistically significant for math and science.

On other outcomes, including teacher ratings of classroom behavior, retention in grade, disciplinary infractions, and attendance, there were generally few overall differences between VPK participants and control children across the years, although school records did show somewhat more disciplinary actions for the violation of school rules for the VPK participants. For special education designations, the differences were not so modest – VPK participants had distinctly higher rates than control children that extended through the 3rd grade year. However, VPK participation allowed children deemed in need of such services to be identified during the pre-k year and, unlike the control children, to already have special education designations when they entered kindergarten. Those early designations persisted through later school years.

The third research question is whether some demographic subgroups of children benefited more from VPK participation than others. Considering the number of combinations of subgroups, outcomes, and school years involved in examining this issue, relatively few differential pre-k effects were found. Children for whom English was not their native language and the overlapping group with low pretest achievement scores experienced the greatest gains on the achievement measures during the pre-k year, but that early advantage did not persist into the later grades. For school and classroom behavior, there was an intriguing pattern of differential pre-k effects on some measures for male and Black children. In the years after kindergarten these children received higher teacher ratings than their control counterparts on a mix of outcomes related to peer interactions, appropriate engagement in classroom activities, and feelings about school, and their school records showed somewhat higher attendance rates.

The generality of the overall findings from the samples of programs and children that provided the data for this study was affirmed in several ways. Pre-k effect estimates for the consented ISS subsample showed no pattern of statistically significant differences from estimates for the remaining unconsented children in the full RCT sample on the outcomes available for both groups. These results provide some assurance that the consent procedure did not select a subgroup of children for whom pre-k effects on outcomes only available for them are unrepresentative of the effects that would be expected for the full RCT sample.

Additionally, the fortuitous availability of a probability sample of VPK programs and participants statewide made it possible to extrapolate the pre-k effects found with the RCT and ISS samples to the statewide VPK population. Though not definitive, this procedure gave no indication that the overall findings of this study would not generalize to the statewide population of VPK programs and the children who participate in them.

6.1. Issues raised by the VPK findings

6.1.1. Program quality

The unfavorable outcomes of the Tennessee program might be explained if it is of distinctively poor quality and, as such, unrepresentative of other state programs that might be assumed to be more effective (Bustamante, Hirsh-Pasek, Vandell, & Golinkoff, 2017). Whether state pre-k programs generally are of high or low quality is an open question, but we know of no evidence that demonstrates that VPK is notably below average. The only attempt to compare quality across states is in the NIEER reports that identify the 10 standards NIEER advocates and which are met by each state program. By those criteria, Tennessee has one of the better programs, meeting 9 of those standards.

The results of classroom observations provide another perspective. Until recently, the most commonly used quality measure for early childhood classrooms has been the *Early Childhood Environmental Rating System-Revised (ECERS-R;* Harms, Clifford, & Cryer, 1998). ECERS-R data were collected by trained observers during daylong mid-year visits to the classrooms of the statewide representative sample of VPK programs drawn for our separate RDD study. The average total ECERS-R score for VPK classrooms was 4.15. Scores that have been reported for other state pre-k programs show variability but by our calculations average 4.35 (Bassok, Fitzpatrick, Greenberg, & Loeb, 2016; Coley, Votruba-Drzal, Collins, & Cook, 2016; Keys et al., 2012; Weiland, Ulvestad, Sachs, & Yoshikawa, 2013). Neither that average nor that for VPK crosses the 4.5 quality threshold espoused by Burchinal et al. (2016), but this comparison does not indicate distinctively lower quality for VPK classrooms.

The search for an evidence-based definition of high quality pre-k that would better predict both short and long-term positive effects is intensifying (see e.g., Sharpe, Davis, & Howard, 2017). However, recent empirical approaches have not produced definitive results. None of the widely used classroom assessments has been linked strongly to child outcomes (Burchinal, 2017; Mashburn, 2016). Nor have teacher qualifications been found to be consistently associated with better child outcomes (e.g., Lin & Magnuson, 2018). Walters' (2015) analysis of the relationship between characteristics of Head Start centers and child outcomes using data from the Head Start Impact study led him to conclude that "important

Please cite this article in press as: Lipsey, M. W., et al. Effects of the Tennessee Prekindergarten Program on children's achievement and behavior through third grade. *Early Childhood Research Quarterly* (2018), https://doi.org/10.1016/j.ecresq.2018.03.005

drivers of successful preschool programs have yet to be identified" (Walters, 2015, p. 99). Recent work by Farran and colleagues has highlighted promising features of classroom interactions predictive of child growth (Farran, Meador, Christopher, Nesbitt, & Bilbrey, 2017), but data on those interactions are expensive to collect and are not included in any current widely used classroom assessments.

The upshot of this fluid research enterprise is that we do not yet have a basis for improving state-funded pre-k programs that is grounded in empirical evidence relating program characteristics to child outcomes. The most direct quality index we have may be the nature and magnitude of the impact a pre-k program actually has on outcomes that are consequential for participating children. This perspective is clouded by uncertainty about which outcomes in fact are most consequential (Bailey et al., 2017). The study reported here demonstrates that VPK has substantial impact on a number of widely used outcome measures during the pre-k year, but longer-term effects were largely null or even negative. Judged by effects on child outcomes, these findings present a mixed picture. Whether other state programs have larger or longer lasting effects is difficult to assess in light of the limited and methodologically problematic research available on other programs described in the introduction to this paper. But if further research demonstrates that they do, it is imperative that researchers also provide detailed information about the practices that led to those outcomes.

6.1.2. The counterfactual

All evaluations of the effects of programs and policies must assess them against a counterfactual condition, meaning in the case of pre-k, what children would experience if they were not in a pre-k program. The small experimental programs of the late 1950s through the early 1970s that have such a high profile in discussions of the benefits expected from public pre-k served primarily poor African American children (Darlington, Royce, Snipper, Murray, & Lazar, 1980) during an era of segregation and immense poverty (Jencks, 1972).

Young children today are different. Between 1998 and 2010 disparities in school readiness between children from lower and higher income families narrowed despite sharp increases in the income gap and segregation in housing and schools (Reardon & Portilla, 2016). Reasons for this surprising finding have to be speculative, but Bassok, Finch, Lee, Reardon, and Waldfogel (2016) document changes in parenting practices that may be relevant. Comparing 1998–2010, they found that parents increasingly structured their children's experiences to focus on learning opportunities such as those that involve computer access, more books in the home, and enrichment activities organized specifically for children. It is especially notable that the socioeconomic gap in these practices narrowed over this period with low-income parents showing stronger increases in their investments in their children than more affluent parents.

The greater contrast between the earlier counterfactual for preschool and the contemporary counterfactual is evident in the effect sizes found in the Perry Preschool and Abecedarian studies, which are three times larger than those found for programs evaluated in the past 15 years (Duncan & Magnuson, 2013). With smaller families (Angier, 2013), exposure to *Sesame Street* (Kearney & Levine, 2015), and access to the Internet even for young, poor children (Rideout & Katz, 2016), early childhood experiences are very different in the 21st century. Contemporary pre-k programs that, as a matter of course, only provide a more concentrated version of what children would otherwise experience are unlikely to lead to the strong effects seen 50 years ago.

6.1.3. Public school involvement in pre-kindergarten programs

The issue of the appropriate vision for early childhood education raises the perennial question of which agency should administer pre-k programs. Departments of health and human services and education have been the primary contenders. As states and the federal government increase funding for pre-k, education is becoming the lead agency and pre-k classrooms are more often located in public elementary schools (94% in Tennessee) than in community-based early childhood centers.

One consequence of this trend is reflected in our finding that more VPK children than control children were identified for special education in contrast to what earlier studies led the field to expect (an expectation central to return-on-investment strategies like Pay for Success; http://www.payforsuccess.org/learn/issues#early-childhood-education). Children who enter the public school system a year before kindergarten not only have a greater chance of being identified for special services prior to kindergarten entry, but their varied developmental progress at that younger age may cause more of them to be identified than would be the case a year later. It is notable in this regard that the overwhelming majority of special education designations in VPK were for speech and language issues, a domain in which development is especially varied for 4-year olds. Once a child has received a special education designation, it is difficult to lose it. While early special education intervention may be a good thing in the long run for the children involved, the expectation that school-based pre-k will lead to rather immediate reductions in special education rates may be unrealistic.

Questions about the consequences of an increasing public school dominance of pre-k programs have existed for years (e.g., Rescorla, Hyson, & Hirsh-Pasek, 1991). Some scholars have argued that a focus on social–emotional development and self-regulation skills is especially important for children of preschool age as a foundation for later academic success (e.g., Bierman, Greenberg, & Abenavoli, 2016). Public schools, however, have increasingly prioritized content and skills related to academic achievement in recent decades with effects that have crept into pre-k and are evident in kindergarten (Bassok, Latham, & Rorem, 2016; Brown, 2009). McCabe and Sipple (2011) characterized the relationship between early childhood education and the public schools as "colliding worlds," and cautioned that the gathering momentum of public school involvement in pre-k education was occurring without a full appreciation of the complexity of providing quality education for young children. If public pre-k is to be offered through the public school system, new fundamental empirical work may be required to identify the appropriate goals and associated instructional practices most important for young children's development in that environment.

6.1.4. Alignment with K-3

Finally, our findings highlight the importance of the K-3 experience. One possible explanation for why the gains children made in VPK did not continue to advantage them afterwards is failure of kindergarten and later teachers to build on the skills those children bring from their pre-k experience (Stipek, Franke, Clements, Farran, & Coburn, 2017). For instance, teachers may teach to the children who need it the most while learning for more advanced children languishes. While this is an empirical question we do not yet have sufficient data to address, explorations of kindergarten teaching suggest that many teachers may be out of touch with the skills children bring to their classrooms. Claessens, Engel, and Curran (2014), for example, found that, rather than focusing on children with the greatest need, kindergarten teachers provided relatively undifferentiated instruction that covered skills many children had already mastered.

Children in VPK classrooms and their counterparts in the control group were eligible for the program because their families were impoverished. VPK participants showed meaningful achievement gains during the pre-k year relative to control children, but after pre-k most then attended high poverty and generally low performing schools. Of concern from our findings is that the achievement

Please cite this article in press as: Lipsey, M. W., et al. Effects of the Tennessee Prekindergarten Program on children's achievement and behavior through third grade. *Early Childhood Research Quarterly* (2018), https://doi.org/10.1016/j.ecresq.2018.03.005

of both the VPK and control children began to decline in 2nd and 3rd grade relative to the national norms for the Woodcock–Johnson tests we used. While Reardon (2011) has rightfully called attention to the widening achievement gap between the rich and the poor, it is important to determine when that actually begins. Our data suggest that children were responsive to their first introduction to formal schooling, whether in pre-k or kindergarten, no matter what their skills upon entry. But their 1st through 3rd grade instructional experiences did not maintain their momentum. It is doubtful that anything done in pre-k can have sustained effects if the gains made there are not supported and extended in the schooling that follows.

7. Conclusion

We are mindful of the limitations of any one study, no matter how well done, and the need for a robust body of research before firm conclusions are drawn. Nonetheless, the inauspicious findings of the current study offer a cautionary tale about expecting too much from state pre-k programs. The fact that the Head Start Impact study – the only other randomized study of a contemporary publicly funded pre-k program – also found few positive effects after the pre-k year adds further cautions (Puma et al., 2012). State-funded pre-k is a popular idea, but for the sake of the children and the promise of pre-k, credible evidence that a rather typical state pre-k program is not accomplishing its goals should provoke some reassessment. It is apparent that the phrase "high-quality pre-k" does not convey enough about what the critical elements of a program should be. Now is the time to pay careful attention to whether the country's youngest and most vulnerable children are well served in the pre-k classroom environments currently operated and to explore innovations with the potential to serve them better.

Acknowledgements

This research was supported by Grant #R305E090009 from the Institute of Education Sciences, U.S. Department of Education and the U.S. Department of Health and Human Services NICHD Grant #R01HD079461-01 for the continuing follow up of the sample. It would not have been possible without the assistance of the Tennessee Department of Education, especially Connie Casha, former Director of the Office of Early Learning; Bobbi Lussier, former Assistant Commissioner of Special Populations; and Robert Taylor, consultant and former Superintendent of Bradley County Schools. The Tennessee Consortium on Research, Evaluation, and Development (now the Tennessee Education Research Alliance) provided essential access to the Tennessee education data system. Special thanks go to colleagues who have served this project well in so many ways: Carol Bilbrey, Project Manager for five years; Kerry Hofer, senior data manager and analyst; Jane Hughart, who stepped in when Dr. Bilbrey retired; Nianbo Dong who served as a key data analyst; Georgine Pion who assisted with the multiple imputations; and the many research assistants who conducted child assessments across Tennessee. We are also grateful for the support of multiple school districts and school administrators throughout Tennessee.

References

Andrews, R. J., Jargowsky, P., & Kuhne, K. (2012). *The effects of Texas's targeted pre-kindergarten program on academic performance (CALDER working paper no. 84)*. Washington, DC: American Institutes for Research. http://caldercenter.org/publications/effects-texas%E2%80%99s-targeted-pre-kindergarten-program-academic-performance

Angier, N. (2013, November 25). *The changing American family*. New York Times. http://www.nytimes.com/2013/11/26/health/families.html?pagewanted=al

Angrist, J. D. (2006). Instrumental variable methods in experimental criminological research: What, where, and how. *Journal of Experimental Criminology*, *2*, 23–44.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 444–455.

Bailey, D., Duncan, G., Odgers, C., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness*, *10*, 7–39. http://dx.doi.org/10.1080/19345747.2016.1232459

Barnett, S., Friedman-Krauss, A., Gomez, R., Weisenfeld, G., Horowitz, M., Kasmin, R., et al. (2017). *The state of preschool 2016: State preschool yearbook*. Rutgers University: The National Institute for Early Education Research.

Bassok, D., Finch, J., Lee, R., Reardon, S., & Waldfogel, J. (2016). Socioeconomic gaps in early childhood experiences: 1998 to 2010. *AERA Open*, *2*, 1–22. http://dx.doi.org/10.1177/2332858416653924

Bassok, D., Fitzpatrick, M., Greenberg, E., & Loeb, S. (2016). Within- and between-sector quality differences in early childhood education and care. *Child Development*, *87*, 1627–1645. http://dx.doi.org/10.1111/cdev.12551

Bassok, D., Latham, S., & Rorem, A. (2016). Is kindergarten the new first grade? *AERA Open*, *1*, 1–31. http://ero.sagepub.com/content/spero/2/1/2332858415616358.full.pdf

Bierman, K. L., Greenberg, M. T., & Abenavoli, R. (2016). *Promoting social and emotional learning in preschool: Programs and practices that work*. Edna Bennet Pierce Prevention Research Center, Pennsylvania State University.

Brown, C. (2009). Pivoting a prekindergarten program off the child or the standard? A case study of integrating the practices of early childhood education into elementary school. *The Elementary School Journal*, *110*, 202–227. http://www.jstor.org/stable/10.1086/605770

Burchinal, M. (2017, April). Challenges in using widely used observational quality ratings. In M. Burchinal (Ed.), *Measuring quality in early childhood education: Issues and promising new instruments. Symposium presented at the biennial conference of the Society for Research in Child Development*.

Burchinal, M., Xue, Y., Auger, A., Tien, H.-C., Mashburn, A., Peisner-Feinberg, E., et al. (2016). Testing for quality thresholds and features in early care and education. In M. Zaslow, & L. Tarullo (Eds.), *Quality thresholds, features, and dosage in early childhood education: Secondary data analyses of child outcomes. Monographs of the Society for Research in Child Development* (pp. 46–63). http://onlinelibrary.wiley.com/doi/10.1111/mono.v81.2/issuetoc

Bustamante, A., Hirsh-Pasek, K., Vandell, D., & Golinkoff, R. (2017, March). *Realizing the promise of high quality early childhood education*. Brookings Institution. https://www.brookings.edu/blog/education-plus-development/2017/03/27/realizing-the-promise-of-high-quality-early-childhood-education/

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.

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. http://dx.doi.org/10.3102/0002831213513634

Cleary, P. D., & Angel, R. (1984). The analysis of relationships involving dichotomous dependent variables. *Journal of Health and Social Behavior*, *25*(3), 334–348.

Coley, R., Votruba-Drzal, E., Collins, M., & Cook, K. (2016). Comparing public, private, and informal preschool programs in a national sample of low-income children. *Early Childhood Research Quarterly*, *36*, 91–105. http://dx.doi.org/10.1016/j.ecresq.2015.11.002

Cooper, D., & Farran, D. C. (1991). *Cooper–Farran Behavioral Rating Scale*. Clinical Psychology Publishing Company, Inc.

Darlington, R., Royce, J., Snipper, A., Murray, H., & Lazar, I. (1980). Preschool programs and later school competence of children from low-income families. *Science*, *208*, 202–204. http://www.jstor.org/stable/1683958

Duncan, G., & Magnuson, K. (2013). Investing in preschool programs. *Journal of Economic Perspectives*, *27*, 109–132. http://dx.doi.org/10.1257/jeb27.2.109

Executive Office of the President of the United States. (2014). *The economics of early childhood investments*. Washington, DC: Author (December) Retrieved from https://www.whitehouse.gov/sites/default/files/docs/early_childhood_report1.pdf.

Farran, D. C., Bilbrey, C., & Lipsey, M. (2003). *Academic and classroom behavior record*. Unpublished scale available from D.C. Farran, Peabody Research Institute, Vanderbilt University, Nashville, TN.

Farran, D. C., Meador, D., Christopher, C., Nesbitt, K., & Bilbrey, L. (2017). Data-driven improvement in prekindergarten classrooms: Report from a partnership in an urban district. *Child Development*, *88*, 1466–1479. http://dx.doi.org/10.1111/cdev.12906

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.

Gennetian, L., Morris, P., Bos, J., & Bloom, H. (2005). Constructing instrumental variables from experimental data to explore how treatments produce effects. In H. Bloom (Ed.), *Learning more from social experiments (2005)*. New York: Russell Sage Foundation.

Gormley, W. T., Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal pre-K on cognitive development. *Developmental Psychology*, *41*, 872–884. http://dx.doi.org/10.1037/0012-1649.41.6.872

Gormley, W. T., Phillips, D., & Anderson, S. (2018). The effects of Tulsa's pre-k program on middle school student performance. *Journal of Policy Analysis and Management*, *37*(1), 63–87.

Grehan, A., Cavalluzzo, L., Gnuschke, J., Hanson, R., Oliver, S., & Vosters, K. (2011). *Participation during the first four years of Tennessee's Voluntary Prekindergarten*

Please cite this article in press as: Lipsey, M. W., et al. Effects of the Tennessee Prekindergarten Program on children's achievement and behavior through third grade. *Early Childhood Research Quarterly* (2018), https://doi.org/10.1016/j.ecresq.2018.03.005

program (Issues & Answers Report, REL 2011, No. 107). Washington, DC: U.S. Department of Education, Institute of Education Sciences, Regional Educational Laboratory Appalachia. Retrieved from http://ies.ed.gov/ncee/edlabs.
Harms, T., Clifford, D., & Cryer, D. (1998). *Early childhood environmental rating scale*. New York: Teachers College Press.
Heckman, J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science*, *312*, 1900–1902. http://www.jstor.org/stable/3846426
Huang, F. L., Invernizzi, M. A., & Drake, E. A. (2012). The differential effects of preschool: Evidence from Virginia. *Early Childhood Research Quarterly*, *27*, 33–45.
Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. New York: Cambridge University Press.
Jackson, R., McCoy, A., Pistorino, C., Wilkinson, A., Burghardt, J., Clark, M., et al. (2007). *National evaluation of early reading first: Final report*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, U.S. Government Printing Office 2007.
Jencks, C. (1972). *Inequality: A reassessment of the effect or family and schooling in America*. New York: Basic Books.
Kearney, M., & Levine, P. (2015). *Early childhood education by MOOC: Lessons from Sesame Street. Working paper 21229*. Cambridge, MA: National Bureau of Economic Research. http://www.nber.org/papers/w21229
Keys, T., Farkas, G., Burchinal, M., Duncan, G., Vandell, D., Li, W., et al. (2012). Preschool center quality and school readiness: Quality effects and variation by demographic and child characteristics. *Child Development*, *84*, 1171–1190. http://dx.doi.org/10.1111/cdev.12048
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. http://dx.doi.org/10.1002/pam.21734
Lin, Y.-C., & Magnuson, K. (2018). Classroom quality and children's academic skills in child care centers: Understanding the role of teacher qualifications. *Early Childhood Research Quarterly*, *42*, 215–227. http://dx.doi.org/10.1016/j.ecresq.2017.10.003
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. (2013). *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. W., Weiland, C., Yoshikawa, H., Wilson, S. J., & Hofer, K. G. (2015). The prekindergarten age-cutoff regression-discontinuity design: Methodological issues and implications for application. *Educational Evaluation and Policy Analysis*, *37*(3), 296–313.
Mashburn, A. (2016). Evaluating the validity of classroom observations in the Head Start Designation Renewal System. *Educational Psychologist*, http://dx.doi.org/10.1080/00461520.2016.1207539
McCabe, L., & Sipple, J. (2011). Colliding worlds: Practical and political tensions of prekindergarten implementation in public schools. *Educational Policy*, *25*, e1–e26. http://dx.doi.org/10.1177/0895904810387415
McFarland, J., Hussar, B., de Brey, C., Snyder, T., Wang, X., Wilkinson-Flicker, S., et al. (2017). *The Condition of Education 2017 (NCES 2017-144)*. Washington, DC: U.S. Department of Education, National Center for Education Statistics. Retrieved from https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2017144.
Minervino, J., & Pianta, R. (2014, September). 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, paper. Statistics and Data Analysis*, 438–2013.
National Institute for Early Education Research (NIEER) (nd). State public preschool quality standards checklist. http://www.megrants.org/programs/201006earlychildhood funders/nieer%20standards.pdf.
Parker, E., Workman, E., & Atchison, B. (2016, January). *50 state review. States pre-k funding for 2015–16 fiscal year: National trends in state preschool funding*. Denver, CO: Education Commission of the States.
Peisner-Feinberg, E. S., Mokrova, I. L., & Anderson, T. L. (2017). Effects of participation in the North Carolina Pre-kindergarten Program at the end of kindergarten: 2015–2016 statewide evaluation. Chapel Hill, NC: University of North Carolina, FPG Child Development Institute.
Phillips, D. A., Lipsey, M. W., Dodge, K. A., Haskins, R., Bassok, D., Burchinal, M. R., et al. (2017). *Puzzling it out: The current state of scientific knowledge on pre-kindergarten effects. A consensus statement*. Washington, DC: The Brookings Institution.
Puma, M., Bell, S., Cook, R., & Heid, C. (2010). *Head Start Impact study. Technical report*. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families.
Puma, M., Bell, S., Cook, R., Heid, C., Broene, P., Jenkins, D., et al. (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.
Reardon, S., & Portilla, X. (2016). Recent trends in income, racial, and ethnic school readiness gaps at kindergarten entry. *AERA Open*, *2*, 1–18. http://dx.doi.org/10.1177/2332858416657343
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.
Rescorla, L., Hyson, M., & Hirsh-Pasek, K. (Eds.). (1991). *Early academics: Challenge or pressure? New directions for child development source book*. San Francisco: Jossey-Bass.
Rideout, V., & Katz, V. (2016). *Opportunity for all? Technology and learning in lower-income families. A report of the families and media project*. New York: The Joan Ganz Cooney Center at Sesame Workshop.
Sharpe, N., Davis, B., & Howard, M. (2017). *Indispensable policies & practices for high-quality pre-k*. Washington, DC: New America Foundation.
Stipek, D., Franke, M., Clements, D., Farran, D., & Coburn, C. (2017). PK -3: What does it mean for instruction? *Social Policy Report*, 30, Number 2 | 2017 ISSN 1075-7031 www.srcd.org/publications/social-policy-report.
TN Comptroller of the Treasury. (2009). *Policy history: Tennessee's pre-kindergarten program*. Offices of Research and Education Accountability. www.tn.gov/comptroller/orea
TNDOE Office of Early Learning. (2014). *List of approved pre-kindergarten curricula 2014-15*. https://www.tn.gov/assets/entities/education/attachments/prek_fact_sheet.pdf
TNDOE. (2013). *Scope of services for 2013–14 Voluntary Pre-K for Tennessee programs*. https://www.tn.gov/assets/entities/education/attachments/prek_scope_of_services.pdf
Tennessee Alliance for Early Education. (2008). *Voluntary Pre-K in Tennessee: Understanding the collaboration model*. http://www.tennessee.gov/assets/entities/education/attachments/prek_understand_collaboration_model.pdf
Walters, C. (2015). Inputs in the production of early childhood human capital: Evidence from Head Start. *American Economic Journal: Applied Economics*, *7*, 76–102. http://dx.doi.org/10.1257/app.20140184
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.
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. http://dx.doi.org/10.1016/j.ecresq.2012.12.002
Wong, V. C., Cook, T. D., Barnett, W. S., & Jung, K. (2008). An effectiveness-based evaluation of five state pre-kindergarten programs. *Journal of Policy Analysis and Management*, *27*, 122–154. http://dx.doi.org/10.1002/pam.20310
Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson tests of cognitive Abilities-III*. Itasca, IL: Riverside.

Please cite this article in press as: Lipsey, M. W., et al. Effects of the Tennessee Prekindergarten Program on children's achievement and behavior through third grade. *Early Childhood Research Quarterly* (2018), https://doi.org/10.1016/j.ecresq.2018.03.005