Running head: FIVE STATE PREKINDERGARTEN STUDY

Effects of Five State Prekindergarten Programs on Early Learning

W. Steven Barnett and Kwanghee Jung

The National Institute for Early Education Research

Vivian Wong and Tom Cook

Northwestern University

Cynthia Lamy

Robin Hood Foundation

October 2007

Abstract

This study estimated the effects of five state-funded preschool education programs on children's learning at the beginning of kindergarten. A regression discontinuity design was applied to programs in Michigan, New Jersey, Oklahoma, South Carolina, and West Virginia. The combined sample included over 5000 children. Receptive vocabulary and print awareness were assessed in all five states. Math abilities were assessed in all except South Carolina. Results are presented from two different sets of analysis. One analysis applied a single model to the entire sample with interactions for each state providing separate estimates for each state. The average effect sizes across these state programs are .18 for receptive vocabulary, .74 for print awareness, and .43 for math. We also present estimates produced by analyzing data on each state independently and selecting the "best" model from 10 alternatives for each outcome and each state. These estimates are more variable and slightly smaller on average. Overall, the evidence indicates that these programs had meaningful effects. These results may not characterize state pre-K programs more generally, as many have lower standards than the state programs in this study. Common elements across the programs in this study are that all or nearly all teachers have a four-year college degree with an early childhood specialization, teacher compensation is comparable to that in the public schools, and class size does not exceed 20 with a full-time aide.

Effects of Five State Prekindergarten Programs on Early Learning

Introduction

State-funded prekindergarten programs have become increasingly common across the country, having been established to some extent in 38 states (Barnett, Hustedt, Hawkinson, & Robin, 2006). The primary goal of these state-funded programs is to improve the learning and development of young children and, thereby, improve their preparation for the increasingly rigorous challenges of kindergarten. Effective preschool education programs lay a foundation for children's subsequent school success, by building knowledge and abilities with an emphasis on language and emergent literacy, but not neglecting other aspects of cognitive development, and attending equally to the development of dispositions, habits, and attitudes, as well as social and emotional development including self-regulation (Frede, 1998). Such an approach is broad and integrated. The need for a broad and integrated approach is recognized by most states, as 28 now have comprehensive standards for their pre-K programs that address the needs of the whole child, and it has long been recognized by Head Start, the nation's major federal pre-K program (Barnett et al., 2006).

Rigorous studies in the United States and abroad have shown the value of high-quality preschool education programs for improving children's short- and long-term success in school and in life (Barnett, 2002; Currie, 2001; Engle et al., 2007). Several studies have found very long-term effects including higher achievement test scores and educational attainment, increased adult productivity, and decreased crime and delinquency (Campbell, Pungello, Miller-Johnson, Burchinal, & Ramey, 2001; Garces, Thomas, & Currie, 2002; Goodman & Sianesi, 2005; Ludwig & Miller, 2007; Raine, Mellingen, Liu, Venables, & Mednick, 2003; Schweinhart et al., 2005; Temple & Reynolds, 2007). However, questions have been raised about whether current

state-funded prekindergarten programs can produce similar effects (Haskins, 1989; Magnuson, Ruhm, & Waldfogel, 2007). Compared to the programs providing the strongest evidence, state prekindergarten programs tend to be shorter in duration (most start at age 4) and less intensive, and to serve more diverse populations. Moreover, there is substantial cross state variation in standards and funding. Most state prekindergarten programs target children who are at elevated risk of school failure, and 27 state programs apply a means test for eligibility (Barnett et al., 2006). Targeted programs have been the most studied (Gilliam & Ripple, 2004; Gilliam & Zigler, 2001). A few states have recently sought to make prekindergarten education available to all 4-year-olds. Less research has been conducted on the impacts of programs for children who are not economically disadvantaged (Blau & Currie, 2006).

As the number of state funded prekindergarten programs grows, it is important to assess how effective they are in improving children's learning and development. However, it has proven difficult to conduct rigorous evaluations of state pre-K programs (Gilliam & Zigler, 2001). One challenge is that the large size and scope of many programs may make it logistically difficult and expensive to obtain a representative sample. Too often evaluations of state programs have been limited to small samples that may or may not generalize to the entire program. Another is that self-selection and administrative selection into programs create the potential for serious selection bias due to unmeasured differences between treatment and comparison groups. More recently, some states have moved toward programs that are universal and (together with Head Start) seek to serve all children. With universal programs it may be practically impossible to find comparable children who do not attend the state preschool program, at least in communities where the program is offered.

Most evaluations of the effects of large-scale public preschool education programs have relied upon statistical models to estimate program effects, adjusting for known and measured differences between children who attended and did not attend those programs (Blau & Currie, 2006; Gilliam & Zigler, 2001). For example, that approach has been followed in two large longitudinal studies, the NICHD study of early child care in the United States (NICHD Early Child Care Research Network, 2002) and the Early Provision of Preschool Education study in England (Sammons et al., 2003). However, such studies are expensive and relatively rare. Most studies of large-scale public preschool programs find it difficult to obtain pre-intervention test scores or to obtain detailed, accurate data on child and family characteristics for use in statistical models. Moreover, even in the best of circumstances it is difficult to rule out the possibility that the characteristics that lead families to choose or not choose state-funded preschool education for their children are not adequately captured by the available data. Thus, it is difficult to eliminate the suspicion that estimates of program effects suffer from selection bias in many instances (Magnuson et al., 2007).

Evidence on the extent to which selection bias is a substantive problem can be obtained by comparing Head Start effects estimates from true experiments with those produced by analyses of data from large-scale surveys. Recently, this has become possible due to a national randomized trial of Head Start, which found positive or null effects on a broad range of measures of cognitive and social-emotional development (Puma et al., 2005). These results are consistent with findings from an earlier, smaller randomized trial (Abbott-Shim, Lambert, & McCarty, 2003). In contrast, regression analyses of data on a national sample of kindergarten children find that Head Start has no effects or even negative effects on cognitive and social and emotional

development (Loeb, Bridges, Bassok, Fuller, & Rumberger, 2007; Magnuson et al., 2007). These results indicate that selection bias is more than a theoretical threat.

In order to address concerns with previous research, we employed a regression-discontinuity design (RDD) to estimate the effects of 5 state prekindergarten programs on children's cognitive development. This method recently has been used to evaluate Oklahoma's universal preschool education program in Tulsa (Gormley, Gayer, Phillips, & Dawson, 2005). The RDD approach explicitly addresses the problem of selection bias and is applicable even when all children attend the program (Cook & Campbell, 1979; Trochim, 1984). Gormley and colleagues (2005) found evidence that the RDD approach reduced selection bias that would have led to underestimation of program effects. In each of our 5 states we used common procedures and measures to assess the impact of the state prekindergarten program for 4-year-olds on learning and development at kindergarten entry. Although substantial samples were obtained in each state, the study was designed so that the state samples could be pooled to increase the study's statistical power. The use of an RDD approach and statewide samples from 5 different states are distinct advantages of this study, but as with every approach there are limitations. One of these is that comparisons of estimates across states are very difficult to interpret. One reason for the difficulty is that the "control" children in our study have access to other kinds of preschool programs. Comparison of this study's estimates to estimates from other studies requires caution, as well.

Methods

The present study was conducted in five states: Michigan, New Jersey, Oklahoma, South Carolina and West Virginia. The prekindergarten programs in Michigan, New Jersey and South Carolina target at-risk children while the programs in Oklahoma and West Virginia are intended

to serve all children. Each state program is unique, but all required licensed teachers with four-year college degrees and certification in early childhoods, with minor exceptions. In Michigan, a small percentage of children attend private programs that do not have to meet the same standards for teacher qualifications as public schools. All programs serve children at age 4, though New Jersey's program serves nearly 80 percent as many children at age 3 as well. In New Jersey, we included only the state's "Abbott District" preschool program, the largest and best funded of that state's three preschool programs. Some states primarily provided services through the public schools, some primarily through private programs. All are well established, though New Jersey's Abbott program could be considered relatively young because it upgraded its standards for teachers and class size with a significant increase in funding beginning in 2002. Some programs are half-day and some full-day, but programs not infrequently find ways to add resources to extend the state-funded day (e.g., Head Start might pay for a half-day and the state for a half-day). Table 1 describes key characteristics of each state program. More detailed descriptions of these state prekindergarten programs are available elsewhere (Barnett et al., 2006).

Research Design

This study employs a regression-discontinuity design (Trochim, 1984). This approach takes advantage of each state program's strict enrollment policy that determines enrollment by the child's date of birth. By relying on this assignment rule, one that is unlikely to related to child and family characteristics, the regression discontinuity design (RDD) seeks to reduce the likelihood of selection bias. Typically, program effects are estimated by comparing the test scores of children who attended a program with the scores of similar children who did not go. Where programs are universal, the problem of finding a "comparable" group of children who did not go to preschool is obvious. Yet, even where programs target only some children, a problem

remains: those who go to preschool are *not* the same as those who do not. Preschool programs that target specific types of children create these differences through their eligibility criteria, but differences also come about because some parents choose to enroll their children and others do not. In sum, children who go to preschool differ from those who do not because programs select children and families select programs.

The RDD approach compares two groups of children who select, and are selected by, a state pre-K program, and to take advantage of the stringent birth date cutoff that states use to define the groups. One way to interpret this design is to view it as similar to a randomized trial near the age cutoff. The RDD creates groups that *at the margin* differ only in that some were born a few days before the age cutoff and others a few days after the cutoff. When these children are about to turn 5 years old the slightly younger children will enter the preschool program and the slightly older children will enter kindergarten having already attended the preschool program. If all of the children are tested at that time, the difference in their scores can provide an unbiased estimate of the preschool program's effect under reasonable circumstances. Obviously, if only children with birthdays one day on either side of the age cutoff were included in a study, the sample size would be unreasonably small. Alternatively, the RDD can be viewed as modeling the relationship between the assignment variable (age) and children's outcomes. The pre-cutoff sample models the relationship prior to treatment. The post-cutoff sample is used to model the relationship after the treatment. This approach can be applied to wider age ranges around the cutoff. However, its validity depends on correctly modeling the relationship. Under either view, it is important that there is minimal misallocation (exceptions to the rule) around the cutoff.

Sample

Typically, samples of children who attend public school programs are drawn from roster lists provided by the school districts. However, we find that many school districts have a difficult time producing valid lists early in the school year, causing delay for research. Since the current research depends on assessing children as early as possible in the academic year, we developed a sampling strategy that required no student lists be provided. Our method was to gather information on the number and location of classrooms across the state universe of state-funded preschool and kindergarten programs and to randomly select enough state-funded preschool classrooms to provide us with the required preschool child sample, assuming approximately four randomly selected children per classroom. We then sampled the same number of kindergarten classrooms as preschool classrooms from each school district with preschool classrooms in the study. Trained research staff visited each sampled program site, selected children into the sample using the class roster and a random number list, and conducted the child assessments as early as possible in the school year.

We initially identified a random sample of 1937 classrooms (approximately half preschool and half kindergarten) in the five states, which would have yielded a sample of over 7600 children. Difficulties obtaining access to some classrooms (for example district refusals to allow participation based on passive consent) and scheduling problems led us to access 1320 classrooms, typically assessing four children per class. Thus, we collected data on 5278 preschool and kindergarten children in the five states. The preschool treatment group includes 2728 children, and the control group includes 2550 children. The sample is quite diverse: 47 percent were White, 25 percent African-American, 21 percent Hispanic, 3 percent Native American, and 2 percent Asian. The sample was roughly evenly split by gender, 48 percent were

boys and 52 percent were girls. Sample size varies from state to state as follows: 871 for Michigan, 2072 for New Jersey, 838 for Oklahoma, 777 for South Carolina, and 720 for West Virginia.

Data Collection Procedures

In each state we worked with a local research partner to train child assessors on issues related to assessing children in school environments, confidentiality, protocol and professional etiquette as well as training specific to the assessment instruments and sampling procedures. Assessors were trained on each assessment and then shadow scored in practice assessments. Site coordinators were responsible for assuring adequate reliability throughout the study. A liaison at each site gathered information on the children's preschool status, usually from existing school records but occasionally from parent report, and was reimbursed $5 per child for obtaining the information.

Children were tested in the fall of the 2004-05 school year. On all measures, children were tested in English or Spanish depending on their strongest language, which was ascertained from the classroom teacher. A very small number of children who did not speak either English or Spanish well enough to be tested were not included in the sample. Assessments were conducted one-on-one in the child's school, and assessments were scheduled to avoid meal, nap and outdoor play times. Testing sessions lasted 20-40 minutes.

Individualized assessments were selected to measure the contributions of the preschool programs to children's learning, with emphasis on skills important for early school success. Criteria for selection of measures included: (1) availability of equivalent tasks in Spanish and English, (2) reliability and validity, particularly pre-literacy skills that are good predictors of

later reading ability; and (3) appropriateness for children ages 3 to 5. Each measure is discussed in detail below.

Measures of Learning

Children's receptive vocabulary was measured by the Peabody Picture Vocabulary Test, 3rd Edition (PPVT-III; Dunn & Dunn, 1997). The PPVT – III is a 204-item test in standard English administered by having children point to one of four pictures shown when given a word to identify. The PPVT-III directly measures vocabulary size and the rank order of item difficulties is highly correlated with the frequency with which words are used. This test is also used as a quick indicator of general cognitive ability, and it correlates reasonably well with other measures of linguistic and cognitive development related to school success. Children tested in Spanish were given the Test de Vocabulario en Imagenes Peabody (TVIP; Dunn, Lugo, Padilla, & Dunn, 1986). The TVIP uses 125 translated items from the PPVT to assess receptive vocabulary acquisition of Spanish-speaking and bilingual students.

The PPVT has been used for many years (over several versions) and substantial information is available on its technical properties. Reliability is good as judged by either split-half reliabilities or test-retest reliabilities. The test is adaptive in that the assessor establishes a floor below which the child is assumed to know all the answers and a ceiling above which the child is assumed to know none of the answers. This is important for avoiding floor and ceiling problems (Rock & Stenner, 2005).

Children's early mathematical skills were measured with the Woodcock-Johnson Tests of Achievement, 3rd Edition (Woodcock, McGrew, & Mather, 2001) Subtest 10 Applied Problems. Spanish-speakers were given the Bateria Woodcock-Munoz Pruebas de Aprovechamiento – Revisado (Woodcock & Munoz, 1990) Prueba 25, Problemas Aplicados. The manuals report

good reliability for the Woodcock-Johnson achievement subtests, and they have been widely and successfully used in studies of the effects of preschool programs including Head Start.

Print Awareness abilities were measured using the Print Awareness subtest of the Preschool Comprehensive Test of Phonological & Print Processing (Pre-CTOPPP; Lonigan, Wagner, Torgeson & Rashotte, 2002). Items measure whether children recognize individual letters and letter-sound correspondences, and whether they differentiate words in print from pictures and other symbols. The percentage of items answered correctly out of 36 total subtest items is reported. The Pre-CTOPPP was designed as a downward extension of the Comprehensive Test of Phonological Processing (CTOPP; Wagner, Torgeson & Rashotte, 1999), which measures phonological sensitivity in elementary school-aged children. Although not yet published, the Pre-CTOPPP has been used with middle-income and low-income samples and includes a Spanish version. As the Pre-CTOPP has only been very recently developed, very little technical information is available about its performance and psychometric properties.

Statistical Analyses

To estimate the effects of the prekindergarten programs on children's test scores we conducted a series of analyses to guard against model misspecification. In this paper we first present analyses conducted pooling data from all of the states to estimate program effects. We begin with complex analyses and gradually pruned the analyses to maximize the efficiency of the estimates. The model accounted for the number of days between birthdates and enrollment cut-off dates for each sample child, gender, ethnicity (classified as African-American, Hispanic, Native American, or White), free lunch status (free-and-reduced-price v. full-price) and whether a child was tested in English or Spanish. Models were estimated with interactions that allowed both intercept and slope to vary by state and we allowed slope to differ pre- and post-treatment

(as these were not significant, the model was simplified). Analyses were conducted using raw scores. All standard errors are clustered by classroom, and STATA (StataCorp, 2005) was used to conduct the regressions.

In these single equation models, the effect of attending the preschool program is estimated at the birth date cut-off for enrollment. A "treatment" variable was defined by assigning all children with birth date after cut-off date with a value of one (treatment) and all other children a value of zero (comparison). The selection variable (the age difference between birth date and cut-off date) was rescaled so that zero-point corresponded to the cut point. Thus, children in the treatment group had positive values, and children in the comparison group had negative values. An interaction term was constructed by multiplying the cut-off dummy variable by the rescaled selection variable. Dummy variables were created for each state and one state served as the "reference" state with state dummy variables interacting with the other variables to allow for separate estimates of intercepts, treatment effects and slopes. We estimated the models with New Jersey as the reference state because of its larger sample size. This provided a model where we could test for differences between each state and New Jersey. After pruning the model, we rotated each state in as the reference state to provide tests of the difference between each state's effect and zero.

As there is no *a priori* expectation that the estimated relationship should be linear, we estimated higher order polynomial forms of the equation, including squared and cubic transformations of the selection variable (the difference between birth date and cut-off date) and its interaction with the cut-off dummy variable). Interactions were used to include higher order terms for each state. We began analyzing third-order (cubic) polynomial regression models and found the coefficients for cubic term (X^3) and its interaction with the cut-off dummy variable

(X^3Z) were not statistically significant. These terms were dropped and the second order model was estimated. When we estimated the second order polynomial, the coefficients for the quadratic terms (X^2) and quadratic interaction terms (X^2Z) were not significant, except for print awareness. Thus, we dropped the quadratic term and its interaction term for the analyses of PPVT-III and Applied Problems scores. For print awareness we focus on the quadratic model results, but report both the linear model results for consistency.

For the regression discontinuity design to be effective, programs must adhere to a fairly strict use of a birth-date cut-off date for program enrollment to determine whether children are enrolled into the kindergarten or prekindergarten program based on their age. Each sample state employed a birth-date cut-off date for program enrollment, which varied by state. Children qualified to attend kindergarten in academic year 2004-05 if they were born before September 1, 1999 in South Carolina, Oklahoma and West Virginia, or before December 1, 1999 in Michigan. They qualified for Pre-K if they were born after those dates but before the same dates in 2000. In New Jersey, the age cut-off for program enrollment varied by school district from September 30 through December 31. Fortunately, there were very few departures from the selection rule.

Our pooled data analyses were "sharp" regression-discontinuity models that employed a total 5071 children in our sample, dropping 207 children (4 percent of the total) whose birth date information appears to be inconsistent with the birth-date cut-off requirement for their programs. The 207 dropped includes both children who appeared to be too young for their grade (n = 60) and children who appeared to be too old for their grade (n = 147). An alternative is to conduct a "fuzzy" regression discontinuity analysis that includes cases where the relationship between children's birthday cut-off dates and their assignment to the treatment or control group is not consistent (Trochim, 1984). "Fuzzy" analyses were conducted using an instrumental variables

(IV) approach that included these children and provided a way to see if their exclusion might have changed the results (Hahn, Todd & Van der Klaauw, 2001). Point estimates were similar to those from the "sharp" analyses. Thus, results of the "sharp" analyses are reported as our primary analyses because they provide more statistical power than the IV analyses and essentially the same effect size estimates.

Our analyses included a measure of whether or not each child qualified for free or reduced lunch under the federal subsidized school lunch program. Unfortunately, this information could not be obtained for 17 percent of the sample. The individual state response rates for this information varied from a particularly low 47 percent in one state to around 90 percent for the others. Analyses including a dummy variable for free and reduced lunch also included a variable indicating when this measure was missing.

One of the key assumptions underlying the regression discontinuity design is that the unobservable characteristics of children do not vary discontinuously around the birth date cutoff. While this is not directly testable, it is possible to see if the observable characteristics vary discontinuously at the cutoff, which is at least suggestive of a possible problem. To test this we re-estimated the regression discontinuity model with minority status and free lunch status as dependent variables. These analyses did not find statistically significant discontinuities at the age cutoff.

In another test of the underlying assumptions of the RDD approach, we repeated our analyses on two subsamples, one limited to children with birthdates within 60 days of the birth date cutoff and the other limited to children with birthdates within 30 days of the birth date cutoff. Analyses of these subsamples tended to produce somewhat smaller estimates of program effects. This is a cause for concern and might indicate problems with the linear functional form.

Thus, we also report results from IV models (thereby including all cases) estimated separately for each state where the functional form was selected based on a graphical review of the data and comparisons of nonparametric analyses, and alternative functional forms employing higher-order polynomials. Wong, Cook, Barnett, & Jung (2007) provide details on estimation of these models and the extensive testing of assumptions including linearity and alternative functional forms that went into model selection. Ten different estimates were produced for each outcome measure in each state varying statistical technique and functional form. Across 14 models (2 outcome measures in 5 states and a third outcome measure in 4 states), the linear functional form was concluded to fit best for 9, the cubic for 3, and the quadratic for 2.

The individual state IV analyses and the pooled data set analyses may be viewed as taking different risks in estimating program effects. The individual IV models provide the best estimate for each state program examined independently using all of the data available for that state. It offers the lowest risk of incorrectly assuming that the functional form is linear. However, this approach increases the risk of error fitting due to sampling variation that might distort estimates. It also has less statistical power due to smaller sample size and increased standard errors associated with the IV approach. The pooled data analyses allow data from all states to inform estimation of effects for each state, increasing statistical power, and giving greater weight to data from states with larger samples. However, if the correct model differs by state, the pooled analysis may raise the risk of missing real differences in the correct functional form and, thereby, producing biased estimates for some states.

It is important to understand how to interpret the RDD results produced by the various approaches. When the response functions are parallel and linear, one can generalize treatment effects across the entire distribution of the assignment variable. When these assumptions do not

hold, only the local average treatment effect at the point of discontinuity is estimated. In that case, treatment effects are estimated only for the sample of children with birthdays near the cutoff.

One way to extend generalization is to have a sampling plan that includes multiple sites that vary in their cut-off points. Researchers can then identify average treatment effects over the achieved range of cut-off values. This is especially useful in education where cutoffs are often site-specific anyway, dependent on local decisions at school, district or state level. As we mentioned earlier, state cut-off dates in our study ranged from September 1st to December 31st, and in New Jersey cutoffs varied across this range by district. In our data set, the cutoff therefore occurs over a range with interpolated points and not a single point.

One advantage of our study is that multiple cutpoints were used by reentering the assignment variable for all states so that they shared a common unique cutoff (0). Thus, instead of estimating local average treatment effect as one would in a regular RDD, we estimated treatment effects over a range of values on the assignment variable.

Results

An overview of the sample is provided by Table 2, which presents descriptive statistics. As can be seen, the sample is highly diverse. The sample has substantial percentages of Hispanic and African-American children. Slightly less than half the sample is white, non-Hispanic. Nearly half the sample is composed of children who do not qualify for free or reduced price lunch; this indicates that they have incomes above 185% of the poverty line. Finally, the two groups are highly similar in their demographic characteristics indicating that the sampling appears to have produced comparable groups.

Tables 3 to 5 present the estimated effects and effect sizes for each outcome measure and include the average effects on children's test scores across the state programs. Statistically significant effects of the state preschool programs were found on all three outcome measures. For each outcome the estimated effect for New Jersey's Abbott program was statistically significant. Only one of the other state estimates was statistically significantly different from New Jersey's estimate with $p < .05$ (West Virginia, for Print Awareness). However, in a number instances the estimated differences would be meaningful in size and would have been significant with a more generous alpha of .10. Thus, in Tables 3 to 5 we report estimated effects and effect sizes for each state program and indicate whether each of these was statistically significantly different from zero. In addition to the pooled data results, we report the estimated effects and effect sizes generated by the "best fit" models estimated individually for each state program. Results for each outcome measure are reviewed in detail below.

Vocabulary

The estimated effects of five state-funded prekindergarten programs on the PPVT-III are reported in Table 3. Applying one common model to data from all states, estimated effects were statistically significant for New Jersey and Oklahoma, which both had effect sizes of about .33. Effect sizes were near zero for Michigan and South Carolina. West Virginia's effect size was intermediate. Averaged across all states the effect size was .18. State specific instrumental variables (IV) models produced a wider range of outcomes, including a disconcerting negative (though not significant) estimate for Michigan. Oklahoma's and New Jersey's IV estimated effect size were statistically significant. The average IV estimated effect size across the 5 states was .14.

Math

The estimated effect of state-funded prekindergarten on children's scores on the Woodcock-Johnson-III Applied Problems subtest (Table 4) was statistically significant for all four states in the pooled analyses. The Applied Problems test was not administered in South Carolina due to resource constraints. As with the PPVT-III, the estimates for other states did not significantly differ from New Jersey's. However, the estimated effect sizes for the other three states were about .50, whereas the estimated effect size for New Jersey was only .19. The average effect size across all four states was .43. Again, the state specific IV estimates were less consistent and only two (Michigan and New Jersey) were statistically significant. The average IV effect size across the 4 states was .29.

Print Awareness

The estimated effects of the five state-funded pre-K programs on children's Print Awareness scores (Table 5) were statistically significant for each state program in the pooled analyses. Print Awareness was the one measure for which the hypothesis of a quadratic functional form was not rejected in the pooled analysis. Modeling the relationship with print awareness poses particular difficulties because the discontinuity is so sharp, many children have little knowledge prior to participating in the pre-K programs and many have mastered much of the knowledge after participating. In the quadratic model, estimated effect sizes ranged from .46 to 1.10 with an average of .74. (The effect sizes in the linear model ranged from .67 to .99 with an average of .84.) The state specific IV models produced effect sizes that were only slightly smaller and ranged from .43 (Oklahoma) to .96 (Michigan) with an average of .70. Only the IV estimated effect for Oklahoma was not statistically significant.

Discussion

This study estimated the effects of five state pre-K programs serving 4-year-olds on several measures of children's early learning relating to language, literacy, and mathematics. This study adds to the evidence that preschool education programs of reasonable quality can produce broad gains in children's learning at kindergarten entry. These kinds of effects we found may be expected to yield greater school success, particularly in reading and math. For example, early print awareness and receptive vocabulary have been found to predict later reading abilities in the early elementary grades (Snow, Burns, & Griffin, 1998). In several longitudinal studies, these kinds of early effects have been predictive of later school success and even positive outcomes for young adults (Campbell, Ramey, Pungello, Sparling, & Miller-Johnson, 2002; Reynolds, Temple, Robertson, & Mann, 2002; Schweinhart, et al., 2005).

The estimated effects found in this study were modest and varied somewhat across the states. The pooled sample estimates tended to be larger, more consistent across states, and more often statistically significant. However, the average effects from the individual state IV models support the same general conclusions. Effects on print awareness were uniformly large (as effects of preschool education go). That the effect on print awareness was good sized across the board may reflect the limited goal and ease with which it is accomplished. Gains in math were more varied and tended to be smaller, though effect sizes of .30 to .40 are still respectable. Gains on the PPVT were smaller and more variable.

In independent IV analyses for each state, response functions differ across states for the same outcome. In general, one would expect response functions to differ if states varied in more ways than expected from sampling error alone. For instance, states may vary in the distribution of children's ages; with very young children, floor effects might be evident and with the oldest,

ceiling effects. Such floor and ceiling effects might generate a cubic response function. States also vary in the population served (percentage who are low income, not fluent in English, or from low-income families) and in the availability of other options for early education available to the population served by state prekindergarten programs. These might vary in such a way as to yield nonlinear functions. Finally, state data collection teams may have differed in the administration of measures, but given that the same assessments, procedures, and data collector training were used in all five states, we believe this is not a major concern. Ultimately, we cannot be definitive as to why response functions varied by state and outcome, and whether the differences were due to substantive differences or sampling error. However, graphical, parametric, and non-parametric evidence indicated some heterogeneity in the response functions, and to simply assume otherwise risks producing biased results. For the most part, these estimates are quite similar to those from the polled analyses, however.

It is noteworthy that effect sizes declined moving from the narrowest measure, print awareness, to the broadest, the PPVT-III, which is variously regarded as a test of receptive vocabulary or a quick test of general cognitive ability. These differences in effect sizes should not be interpreted as meaning that more was learned in the domain of literacy than in the domain of math, while the least progress was made in language. Instead, this may be seen as the operation of a general principle that it is easier to produce large effect sizes the more narrowly defined the outcome measure. In this case, print awareness is quite narrow—there are only 26 letters to be learned. Math is broader, but the range of mathematical abilities and knowledge assessed for children ages 4 to 6 is still fairly narrow on the Applied Problems scale. Vocabulary and the conceptual knowledge tested by the PPVT-III seems likely to be the broadest domain assessed in this study. It is entirely possible that amount of learning that state prekindergarten

produced in language and general cognitive abilities was quite comparable to the amount produced in print awareness and mathematics despite the differences in effect sizes.

This study's results are broadly consistent with findings from other studies of public preschool education programs that employed strong research designs. Estimated effects on the Applied Problems scale are similar to those found by Gormley and colleagues (2005) in their Tulsa study. Effect sizes for their Letter-Word Identification and our Print Awareness tests are quite similar, as well. Similar results were obtained in two other quasi-experimental studies of state pre-K programs with relatively strong strategies for creating comparison groups (Frede & Barnett, 1992; Irvine, Horan, Flint, Kukuk, & Hick, 1982). The PPVT effect size is smaller than the effect size at kindergarten entry in the Chicago Child Parent Center study, which ranged from .46 to .63 depending on the analytical approach (Reynolds & Temple, 1995). However, the Chicago study employed a composite achievement measure that seems more comparable to an average of the Print Awareness and Applied Problems effects, and this average effect size is roughly equal to the Chicago effect size.

Comparisons to other studies are complicated by differences in procedures as well as by the extent to which other preschool education programs were available to the comparison group. The most effective model programs appear to have produced considerably larger effects than we found in this study. This seems consistent with differences in dosage (e.g., smaller classes, two or more years, longer days) and not just due to a lack of access to preschool programs by control groups in earlier studies (Barnett, 1998). Average effect sizes are larger than those found for prekindergarten using national survey data (Magnuson et al., 2007); this could be an artifact of research design or due to real differences between the 5 state programs we studied and prekindergarten programs generally. Our study also appears to have produced larger average

estimated effects than the national randomized trial of Head Start, but several issues make the comparison less than straightforward. The Head Start study has crossover by treatment and control groups as well as participation by control group members in other programs (including state prekindergarten) that tend to result in underestimation of Head Start's effects (Puma et al., 2005). The populations served in the 5 state programs differ from the Head Start population, and the programs available to the control children are likely to differ between the studies. This makes it difficult to learn much from comparing our results to those for Head Start. If we really wish to learn how changes in Head Start that made it more like these 5 state programs (such as increased teacher qualifications and compensation) affect child outcomes, new studies will be required that are specifically designed to address that question.

When interpreting the results of this study, it should be kept in mind that our approach estimated the effects of providing state pre-K for children many of whom have other opportunities for preschool education including Head Start, local school programs, and private programs offered by a wide range of for-profit and non-profit organizations. These other programs may have applied different birth date cut-offs for entry or had no relevant age requirement at all. Thus, the study does *not* estimate the effects of these 5 state-funded pre-K programs relative to no-program. Instead, it estimates the effects of these programs relative to other available alternatives. State policy makers often regard this as the most relevant question because they want to know the benefit from making state dollars available for such programs. However, this complicates the comparisons of estimates within this study and with the results of other studies. Each state's estimated effects depend not only on the contributions of their state-funded pre-K program to learning and development, but on the contributions of the other programs available to children in each state. Access to other preschool programs, regulations

and standards for child care, and the general quality of preschool programs varies among states due to state policies, family incomes, and other factors (Barnett & Yarosz, 2004; Blau & Currie, 2006).

Despite our caution about the difficulties of interpreting and comparing results, the variations in effect size estimates are disconcerting, particularly for the IV estimates. For example, Michigan's program yielded the largest estimated effect for math gains, a larger than average effect for print awareness and the smallest gain on the PPVT. Indeed, the IV estimate for the PPVT is inexplicably negative. Furthermore, given the traditional emphasis of preschool education programs on language and literacy, it would be expected that that any program that excelled in mathematics education would excel in those other areas. Possibly, the Michigan results reflect the effects of fairly high levels of participation in other programs by children who did not yet enter state prekindergarten. These other programs might be relatively strong with respect to language and general cognitive development, but weaker in literacy and math education. Alternatively, Michigan (followed by South Carolina) offers the fewest hours of preschool education per week, and the sheer number of hours of exposure may matter most for language development. However, these are just speculations on our part as we have no direct data on the quality and practices of either the state prekindergarten programs or the alternatives in which children participated. Future studies could benefit from this kind of information.

Comparisons across states also are complicated by differences in program eligibility criteria that create differences in the population served by state-funded pre-K. Oklahoma and West Virginia seek to offer preschool education to all children regardless of income, and the average child served was average as indicated by standard scores of around 100 on the PPVT-III at pre-K entry. The other three state programs in this study targeted children from primarily

lower income families, and the average child served scored from somewhat below (96 in Michigan) to far below (87 in New Jersey) on the PPVT-III. Differences in estimated effects between "universal" and "targeted" programs could reflect differences in responsiveness to the program for children from different family backgrounds or differences in access to other services for children from different family backgrounds. However, no clear pattern of differences in estimated effects emerges from a comparison of Oklahoma and West Virginia to the other state programs in this study.

A notable limitation of our study is the absence of a measure of social and emotional development. If future RDD studies are to incorporate such measures, they will require instruments that do not rate children relative to expectations for their age cohort, which is the most common approach. To date, randomized trials of preschool education programs for 3- and 4-year-olds have found positive effects. Studies relying on natural variation in program participation in large-scale surveys have found negative effects on social and emotional development of children, including negative effects for Head Start (Loeb et al., 2007; Magnuson et al., 2007). These results would appear to be contradicted by the findings of the national randomized trial of Head Start, which raises questions about the validity of the estimates for other types of preschool programs. Other non-experimental studies have raised concerns about the potential for modest negative effects on social and emotional development from early childhood programs more generally (Belsky et al., 2007). Given this mixed picture, further research on social and emotional outcomes is warranted with methods particularly attuned to the avoidance of selection bias.

Overall, the results of this study add to the evidence that high quality public preschool education can improve learning and development on a large scale for both targeted and general

populations. Although these results cannot be safely extrapolated to state programs with weaker standards, these states offer models that others could follow. As noted earlier, effects were similar in size to those found in the Chicago-Child Parent Centers study. Temple and Reynolds (2007) have provided a cost-benefit analysis of the Child Parent Centers based on follow-up data through age 21, and they found that benefits far exceeded costs. The estimated benefits are so large, that if one year of participation in these state programs produced even 10% of the estimated benefits of the Chicago program, they would still be likely to pass a cost-benefit test. Thus, our study adds to the evidence that high-quality public preschool education programs can be sound investments from a purely economic perspective.

The strength of the estimated effects varied by outcome measure, state, and analytical method. The limitations of this study, including the small number of states, severely limit our ability to make sense of these variations. Effect sizes appear to be influenced by the breadth of the measure so care must be taken in judging the importance of an outcome from the effect size alone. Further studies may yield greater understanding of the effects of variations in program and populations served. There is some indication that if future studies employing the RDD had larger sample sizes from each state some of the apparent variation might be reduced. Results were more consistent for New Jersey's Abbott prekindergarten program with a sample of 2072 children than for the other state programs where sample size ran from 720 to 871. More evidence of the value of a larger sample is provided by comparing our estimates for Oklahoma with those from the earlier study of Tulsa with its sample of over 3000 children (Gormley et al., 2005).

Although the RDD approach has proven useful, its limitations are such that the use of multiple methods will continue to be important in the evaluation of state prekindergarten

programs. The strongest approach to obtaining unbiased estimates remains the randomized trial, which also provides substantially greater statistical power for a given sample size (Bloom, 2005; Cook, 2002; Wong et al., 2007). In addition, randomized trials incorporating variations in program design could provide a clearer guide to what program elements make how much of a difference for which populations. Such studies might be done most effectively randomizing groups rather than individuals (St. Pierre & Rossi, 2006). Randomized trials and other quasi-experimental designs also provide a basis for longitudinal follow-ups that can estimate longer-term effects of prekindergarten. Follow-up of the RDD samples for another year would only provide a basis for estimating of the effects of kindergarten, not the lasting effect of prekindergarten. RDD studies might be used with other quasi-experimental designs employing local nonequivalent comparison groups. In some circumstances the RDD might provide assurance that relatively uncomplicated analytical procedures yielded little bias. In others, the RDD results might be used to refine matching techniques and choice of analytical procedures to minimize selection bias. What seems certain is that the field needs to move beyond the simple question of whether public prekindergarten programs work to the questions of what works best for whom, using multiple approaches that together support the development of better programs.

Table 1

Description of State Prekindergarten Programs Studied

State	Year established	Number served by child age	Percent of 4's enrolled	Minimum hours per week	Staff/ child ratio	Max. class size
Michigan	1985	24,729 age 4	19%	10	2:16	18
New Jersey Abbott *	1998— upgraded in 2002	21,286 age 4 16,725 age 3	79%	30	2:15	15
Oklahoma	1990 – universal in 1998	30,180 age 4	65%	(Varies) 12.5-30	2:20	20
South Carolina	1984	17,821 age 4 740 age 3	32%	12.5	2:20	20
West Virginia	1983 – universal by 2010	6,541 age 4 1,370 age 3	33%	(Varies) 12	2:10	20

* New Jersey's Abbott districts include about ¼ of the state's children.

Table 2

Children's Demographics and Scores by Group

Demographics	No Pre-K	Pre-K
Ethnicity		
African American	27%	24%
White	47%	48%
Hispanic	20%	23%
American Indian	3%	3%
Asian	2%	2%
Other	2%	2%
Free or reduced lunch	55%	55%
Gender		
Boys	48%	49%
Girls	52%	51%
Home language English	83%	84%
Tested in Spanish	3%	2%
Scores		
Receptive language		
Raw score	49.02 (18.87)	65.68 (18.58)
Standard score	92.16 (15.38)	94.10 (14.37)
Math		
Raw scores	10.66 (4.07)	15.31 (4.31)
Standard score	97.08 (14.15)	95.89 (12.94)
Print Awareness (% correct)	43.48 (25.83	78.70 (21.47)
Phonological Awareness (% correct)	67.94 (23.70)	77.72 (22.44)
Sample Size (N)	2550	2728

Table 3

Estimated Effects on PVVT Raw Scores (Vocabulary)

	Single Model Pooled Sample (*n*=5031)		Model Varies by State (*n* Varies)		
State	Effect	Effect Size	Effect	Effect Size	*n*
Michigan	0.46	0.03	-2.75	-0.16	863
New Jersey	5.83***	0.34***	6.10*	0.36*	2062
Oklahoma	5.57*	0.32*	5.12*	0.29*	827
South Carolina	0.83	0.05	0.80	0.05	768
West Virginia	3.16	0.18	2.42	0.14	713
Average	3.17	0.18	2.34	0.14	

* p<.05. ** p<.01 *** p <.001 tested separately. No state estimates are significantly different from NJ in pooled analysis.

Table 4

Estimated Effects on Woodcock-Johnson Applied Problems Raw Score (Math)

	Single Analysis Pooled (*n*=4178)		Model Varies by State (*n* Varies)		
State	Effect	Effect Size	Effect	Effect Size	*n*
Michigan	2.00***	0.51***	1.82*	0.47*	865
New Jersey	0.72*	0.19*	0.87*	0.23*	2030
Oklahoma	1.93**	0.49**	1.36	0.35	835
West Virginia	2.01**	0.52**	0.44	0.11	641
Average	1.67	0.43	1.12	0.29	

* p<.05. ** p<.01 *** p <.001 tested separately. No state estimates are significantly different from NJ in pooled analysis. Applied Problems test was not administered in SC.

Table 5

Estimated Effects on CTOPPS Print Awareness (Percent Correct)

	Single Model Pooled Sample (*n*=4880)				Model Varies by State (*n* Varies)		
	Linear		Quadratic				
	Effect	Effect Size	Effect	Effect Size	Effect	Effect Size	*n*
Michigan	25.13***	0.99***	20.07***	0.78	22.14*	0.96*	851
New Jersey	17.07***	0.67***	11.95**	0.46	13.02*	0.50*	1932
Oklahoma	19.60***	0.77***	13.98**	0.54	11.46	0.43	829
S. Carolina	20.20***	0.80***	20.92***	0.81	21.01*	0.79*	757
W. Virginia	24.26***	0.95***	28.43***	1.10	20.15*	0.83*	700
Average	21.25	0.84	19.07	0.74	17.56	0.70	

* p<.05. ** p<.01 *** p <.001 tested separately. State estimates are not significantly different from NJ in Linear model. WV estimate is significantly different from NJ (p < .01) in quadratic model.

References

Abbott-Shim, M., Lambert, R., & McCarty, F. (2003). A comparison of school readiness outcomes for children randomly assigned to a Head Start program and program's waiting list. *Journal of Education for Students Placed at Risk 8*(2), 191-214.

Barnett, W. S. (1998). Long-term effects on cognitive development and school success. In W. S. Barnett & S. S. Boocock (Eds.), *Early care and education for children in poverty* (pp. 11-44). Albany, NY: SUNY Press.

Barnett, W. S. (2002). Early childhood education. In A. Molnar (Ed.), *School reform proposals: The research evidence* (pp.1-26). Greenwich, CT: Information Age Publishing.

Barnett, W. S., Hustedt, J. T., Hawkinson, L., & Robin, K. B. (2006). *The state of preschool: 2006 state preschool yearbook*. New Brunswick, NJ: National Institute for Early Education Research.

Barnett, W. S., & Yarosz, D. J. (2004). Who goes to preschool and why does it matter? *Preschool Policy Matters*, *8*. New Brunswick, NJ: NIEER.

Belsky, J., Vandell, D. L., Burchinal, M., Clarke-Stewart, K. A., McCartney, K., & Owen, M. T. (2007). Are there long-term effects of early child care? *Child Development, 78*(2), 681-701.

Blau, D., & Currie, J. (2006). Pre-school, day care, and after-school care: Who's minding the kids? In E. Hanushek & F. Welch (Eds.), *Handbook of the economics of education, Volume 2* (pp. 1163-1277). New York, NY: North- Holland Publishing Co.

Bloom, H. S. (Ed.) (2005). *Learning more from social experiments: Evolving analytic approaches.* New York, NY: Russell Sage Foundation.

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

Campbell, F. A., Ramey, C. T., Pungello, E. P., Sparling, J., & Miller-Johnson, S. (2002). Early childhood education: Young adult outcomes from the Abecedarian Project. *Applied Developmental Science, 6,* 42-57.

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis, 24*(3), 175-199.

Cook, T. D., & Campbell, D. T. (1979). *Quasi experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin.

Currie, J. (2001). Early childhood education programs. *Journal of Economic Perspectives, 15*(2), 213-238.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test-Third Edition (PPVT-III).* Circle Pines, MN: AGS Publishing.

Dunn, L., Lugo, D., Padilla, E., & Dunn, L. (1986). *Test de Vocabulario en Imagenes Peabody.* Circle Pines, MN: American Guidance Service.

Engle, P. L., Black, M. M., Behrman, J. R., Cabral de Mello, M., Gertler, P. J., Kapiriri, L., et al. (2007). Child development in developing countries 3: Strategies to avoid the loss of developmental potential in more than 200 million children in the developing world. *The Lancet, 369*(9557), 229-242.

Frede, E. C. (1998). Preschool program quality in programs for children in poverty. In W. S. Barnett & S. S. Boocock (Eds.), *Early care and education for children in poverty: Promises, programs, and long-term results* (pp. 77-98). Albany, NY: SUNY Press.

Frede, E., & Barnett, W. S. (1992). Developmentally appropriate public school preschool: A study of implementation of the High/Scope curriculum and its effects on disadvantaged children's skills at first grade. *Early Childhood Research Quarterly, 7,* 483-499.

Garces, E., Thomas, D., & Currie, J. (2002). Longer term effects of Head Start. *The American Economic Review*, *92*(4), 999–1012.

Gilliam, W. S., & Ripple, C. H. (2004). What can be learned from state-funded prekindergarten initiatives? A data-based approach to the Head Start devolution debate. In E. Zigler & S. Styfco (Eds.), *The Head Start debates* (pp. 477-497). Baltimore, MD: Paul H. Brookes Publishing Co.

Gilliam, W. S., & Zigler, E. F. (2001). A critical meta-analysis of all impact evaluations of all state-funded preschool from 1977 to 1998: Implications for policy, service delivery, and program evaluation. *Early Childhood Research Quarterly, 15,* 441-473.

Goodman, A., & Sianesi, B. (2005). Early education and children's outcomes: How long do the impacts last? *Fiscal Studies, 4,* 513-548.

Gormley, W. T., Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal pre-k on cognitive development. *Developmental Psychology, 41*, 872–884.

Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica, 69*(1), 201-209.

Haskins, R. (1989). Beyond metaphor: The efficacy of early childhood education. *American Psychologist, 44,* 274-82.

Irvine, D. J., Horan, M. D., Flint, D. L., Kukuk, S. E., & Hick, T. L. (1982). Evidence supporting comprehensive early childhood education for disadvantaged children. *Annals of the American Academy of Political and Social Science, 461*(May), 74-80.

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

Lonigan, C., Wagner, R., Torgeson, J., & Rashotte, C. (2002). *Preschool Comprehensive Test of Phonological & Print Processing (Pre-CTOPPP)*. Department of Psychology, Florida State University.

Ludwig, J., & Miller, D. L. (2007) Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *Quarterly Journal of Economics, 122*(1), 159-208.

Magnuson, K. A., Ruhm, C., & Waldfogel, J. (2007). Does prekindergarten improve school preparation and performance? *Economics of Education Review, 26,* 33-51.

NICHD Early Child Care Research Network (2002). Early child care and children's development prior to school entry: Results from the NICHD study of early child care. *American Educational Research Journal, 39(1),* 133-164.

Puma, M., Bell, S., Cook, R., Heid, C., Lopez, M., Zill, N., et al. (2005). *Head Start impact study: First year findings*. Washington, DC: US Department of Health and Human Services, Administration for Children and Families.

Raine, A., Mellingen, K., Liu, J., Venables, P., & Mednick, S. A. (2003). Effects of environmental enrichment at ages 3-5 years on schizotypal personality and antisocial behavior at ages 17 and 23 years. *American Journal of Psychiatry, 160(9),* 1627-1635.

Reynolds, A. J., & Temple, J. A. (1995). Quasi-experimental estimates of the effects of a preschool intervention: Psychometric and econometric comparisons. *Evaluation Review, 19*, 347-373.

Reynolds, A. J., Temple, J. A., Robertson, D. L., & Mann, E. A. (2002). Age 21 cost–benefit analysis of the Title I Chicago Child–Parent Centers. *Educational Evaluation and Policy Analysis, 24*, 267–303.

Rock, D. A., & Stenner, J. A. (2005). Assessment issues in the testing of children at school entry. *Future of Children*, *15*(1), 15-34.

Sammons, P., Sylva, K., Melhuish, E., Siraj-Blatchford, I., Taggart, B., & Elliot, K. (2003). *Measuring the impact of preschool on children's social/behavioural development over the preschool period.* London: Institute of Education, University of London.

Schweinhart, L. J., Montie, J., Xiang, Z., Barnett, W. S., Belfield, C. R., & Nores, M. (2005). *Lifetime effects: The High/Scope Perry Preschool study through age 40* (Monographs of the High/Scope Educational Research Foundation, 14). Ypsilanti, MI: High/Scope Educational Research Foundation.

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

St. Pierre, R. G., & Rossi, P. H. (2006). Randomize groups, not individuals: A strategy for improving early childhood programs. *Evaluation Review, 30*, 656-685.

StataCorp (2005). *Stata Statistical Software: Release 9.0*. College Station, TX: StataCorp.

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

Trochim, W. M. K. (1984). *Research design for program evaluation: The regression-discontinuity approach*. Beverly Hills, CA: Sage Publications.

Wagner, R., Torgeson, J., & Rashotte, C. (1999). *Comprehensive Test of Phonological Processing (CTOPP).* Austin, TX: Pro-Ed.

Wong, V. C., Cook, T. D., Barnett, W. S., & Jung, K. (2007). An effectiveness-based evaluation of five state prekindergarten programs using regression-discontinuity. Retrieved October 3, 2007, from http://nieer.org/resources/research/EvaluationFiveStates.pdf.

Woodcock, R. W., & Munoz, A. F. (1990). *Bateria Woodcock-Munoz Pruebas de Aprovechamiento – Revisados.* Itasca, IL: Riverside Publishing.

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