

**ABBOTT PRESCHOOL PROGRAM LONGITUDINAL EFFECTS STUDY
(APPLES) YEAR ONE FINDINGS**

A paper presented at the **National Invitational Conference of the
Early Childhood Research Collaborative**
**University of Minnesota Center for Early Education and Development
and the Federal Reserve Bank of Minneapolis**

Human Capital Conference Series on Early Childhood Development
Critical Issues in Cost Effectiveness in Children's First Decade
Friday, December 7 and Saturday, December 8, 2007
Federal Reserve Bank of Minneapolis



APPLES

BY *Norah*

Ellen Frede, Ph.D.
National Institute for Early Education Research and
The College of New Jersey

W. Steven Barnett, Ph.D.
National Institute for Early Education Research

Kwanghee Jung, Ph.D.
National Institute for Early Education Research

Cynthia Esposito Lamy, Ed.D.
Robin Hood Foundation

Alexandra Figueras, M.S.
National Institute for Early Education Research

Acknowledgements

The research reported in this document was conducted under a Memorandum of Agreement as part of the Early Learning Improvement Consortium (ELIC) with the New Jersey Department of Education (NJ DOE) and with partial funding from The Pew Charitable Trusts. The conclusions are those of the authors and do not necessarily represent the views of the funding agencies.

The authors wish to acknowledge the support and assistance of the other members of the Early Learning Improvement Consortium: Dr. Ellen Wolock, New Jersey Department of Education, and Drs. Holly Seplocha and Janis Strasser, William Paterson State University. We are grateful to Dr. Jacqueline Jones, Assistant Commissioner, Division of Early Childhood Education, NJ DOE for comments on an earlier draft. We wish to express our appreciation to Dr. Thomas Cook and Vivien Wong for advice on analysis of the regression discontinuity design data. Other NIEER staff members were instrumental in data collection; in particular, we thank Amanda Colon, Marilyn Quintana and Jessica Thomas for coordinating training and other project assistance. Data collection was also ably coordinated at William Paterson University by Mary DeBlasio and at the College of New Jersey by Lisa Smith. We appreciate their efficiency and assistance.

Most important, we thank the children, parents, teachers, center directors, principals, early childhood education supervisors and all other educators who have graciously assisted us in this critical data collection and analysis. Without their assistance the research could not have been conducted.

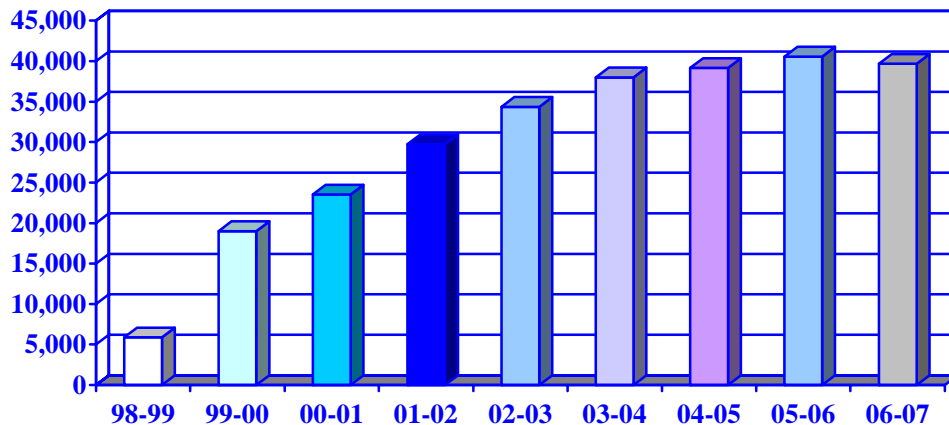
Introduction

This study investigates the educational effects of state funded prekindergarten education for children at ages three and four that came about as a result of a unique situation. As part of the landmark New Jersey Supreme Court school-funding case, *Abbott v. Burke*, the Court established the Abbott Preschool Program. Beginning in the 1999-2000 school year, 3- and 4- year old children in the highest poverty districts in the state were able to receive a high-quality preschool education that would prepare them to enter school with the knowledge and skills necessary to meet the New Jersey Preschool Teaching and Learning Expectations: Standards of Quality (NJ Department of Education, 2004b) and the Kindergarten New Jersey Core Curriculum Content Standards (NJDOE, 2004a). Through a Department of Education (DOE) and Department of Human Services (DHS) partnership, Abbott preschool classrooms combine a DOE-funded six-hour, 180-day component with a DHS-funded wrap-around program that provides daily before- and after-care and summer programs. In total, the full-day, full-year program is available 10 hours per day, 245 days a year.

Enrollment in the Abbott preschool program has increased dramatically since its inception in 1999. During the 2005-2006 school year, the seventh year of Abbott preschool implementation, the 31 Abbott districts served more than 40,500 3- and 4-year-old children in preschool – 78 percent of a possible 52,160 children. The enrollment for the 2006-2007 school year is more than 39,678 children with a DOE budget of almost \$500 million. Through contracts with the school districts, private child care providers and Head Start agencies, in addition to public schools, offer Abbott Preschool: 37 percent

of children are served in district-run classrooms, 7 percent are served in Head Start classrooms, and 56 percent are in private provider classrooms.

Figure 1. Abbott Preschool Program Enrollment 1998-2007



The Court established some basic program standards that included a maximum class size of 15, certified teachers with early childhood expertise, assistant teachers in every classroom, comprehensive services and a developmentally appropriate curriculum designed to meet learning standards. To ensure high quality and consistency for children across auspice and district and to assist administrators and staff who may have been inadequately prepared in early childhood education, more detailed operational standards were developed (Abbott Preschool Program Implementation Guidelines; Office of Early Childhood Education, NJDOE, 2002, revised 2005). They were also necessary because the Court made clear that funding through an across-the-state per pupil formula would not be adequate and that budget decisions must be based on district contexts.

Since 2002, the NJ DOE has implemented an assessment system for the New Jersey Abbott Preschool Program (see Frede, 2005 for details on this system). To measure and assess progress statewide, the DOE formed the Early Learning Improvement

Consortium (ELIC) by bringing together a group of the state's top early childhood education faculty. Drawing on research previously conducted by the Center for Early Education Research (Barnett, Tarr, Esposito Lamy and Frede, 2002), ELIC is responsible for collecting and reporting on data on children and classrooms. Every fall from 2002 through 2005, assessments of kindergartners' skills were conducted to measure progress toward preparing children to succeed in school. In addition, members of ELIC conducted classroom observations on a random sample of Abbott preschool classrooms to measure progress in program quality. Findings have been reported yearly (Frede et al, 2004; Lamy et al, 2005).

In the 2004-2005 school year, ELIC reported that classroom quality scores had reached acceptable levels, and children were entering kindergarten with language and literacy skills closer to the national average than in prior years (Frede, et al, 2004; Lamy, et al, 2005). Given these trends, an evaluation seemed warranted to more precisely estimate the learning gains from the Abbott prekindergarten program and the extent to which gains persist into elementary school. This paper presents the methods and results of that evaluation through the end of the kindergarten year.

The evaluation was conducted in such a way as to build on the previous annual descriptive studies and provides continuity with respect to measures and the sample. However, it moves beyond the previous Abbott preschool program studies by using multiple approaches to the evaluation of the preschool programs and their impacts on children. These mixed methods include: assessing classroom quality with respect to children's experiences and teacher practices; estimating the immediate impact of the program at age four using a regression-discontinuity design (RDD). The RDD approach

explicitly addresses the problem of selection bias and is applicable even if all of the children in a district attend the preschool program (Cook & Campbell, 1979; Trochim, 1984). However, the RDD approach cannot be used to estimate the effects of the program beyond kindergarten entry, nor could it provide estimates of the effects of one v. two years of the program. Thus, we employed a second approach in which we compared three groups of children: those who had not attended the Abbott preschool program, those who had one year of Abbott preschool education, and those who had two years of Abbott preschool education. These children can be followed and compared from kindergarten through elementary school. This second approach to estimating program effects is more likely to suffer from selection bias. However, we can measure the direction and size of such bias by comparing the second set of estimates to those from the RDD approach.

Abbott Preschool Program Quality

Structured classroom observations have been conducted since the inception of the Abbott program in 1999 through 2006. The Center for Early Education Research at Rutgers University (the predecessor of NIEER) measured classroom quality in a subsample of 19 Abbott districts in 1999 through 2001. Thus, change in aspects of classroom quality since before 2002 has been measured across this subsample. Beginning in 2003, ELIC administered observations annually in all Abbott districts. Trained data collectors observed in randomly selected preschool classrooms using structured classroom observation instruments that assess materials, the environment, and teacher-child interactions. Observers typically were advanced undergraduates, graduate students, or former teachers, usually with experience teaching at the preschool level. Each observer was shadow scored and reached an 80% inter-rater reliability rate before

qualifying to conduct observations for the study. Shadow scoring was repeated every six weeks to ensure that observer reliability did not drift over time. Each classroom was observed once during winter or spring of the year for three to four hours. In 2005-2006, the sample consisted of 316 classrooms that proportionately represent the population with respect to auspice (104 public school, 176 private, and 25 Head Start). Each of the three measures is described below.

Early Childhood Environment Rating Scale – Revised (ECERS-R). Overall quality was assessed using the ECERS-R (Harms, Clifford & Cryer, 2005). This measure has been used extensively in the field and has well-established validity and reliability. Internal consistency as measured by Cronbach's alpha is reported by the authors to be adequate (.81 to .91) and was .90 in this study. Classroom quality is rated on a 7-point Likert scale ranging from inadequate (1) to excellent (7). The seven ECERS-R subscales are: Space and Furnishings, Personal Care Routines, Language-Reasoning, Activities, Interaction, Program Structure, and Parents and Staff. Average subscale scores are calculated, as well as a total scale score averaged across all 43 items in the scale.

Supports for Early Literacy Assessment (SELA). The extent to which the classroom environment supports children's literacy development is measured by the SELA (Smith, Davidson & Weisenfeld, 2001). This measure was revised for this study by the deletion of 4 items that overlap with the ECERS-R. The revised measure includes 16 items each rated on a scale from 1 (low support) to 5 (high support). Six subscales are: The Literate Environment, Language Development, Knowledge of Print/Book Concepts, Phonological Awareness, Letters and Words, and Parent Involvement. Internal consistency as measured by Cronbach's alpha on the current sample was good at .87.

Preschool Classroom Mathematics Inventory (PCMI). Classroom support for the development of children's early mathematical skills was measured using the PCMI (Frede, Weber, Hornbeck, Stevenson-Boyd & Colon, 2005). This tool measures the materials and strategies used in the classroom to support children's early mathematical concept development, including counting, comparing, estimating, recognizing number symbols, classifying, seriating, geometric shapes, and spatial relations. The standards of the National Council of Teachers of Mathematics and the National Association for the Education of Young Children (2002) inform the measure, which is comprised of 11 items on a 5-point scale, from 1 (low support) to 5 (high support). It has two subscales: Materials and Numeracy, and Other Mathematical Concepts. Internal consistency among the test items as measured by Cronbach's alpha was good at .86. The PCMI has been found to predict child progress on a standardized math assessment (Frede, Lamy and Boyd, 2005).

To evaluate the extent to which the Abbott preschool program might be expected to produce substantial gains in children's learning and development, we examined means and the distribution of scores across classrooms in 2005-2006. In addition, we compared these results to those obtained in 1999-2000 when quality was essentially what it had been before implementation of the Court order. Taken as a whole, the advances in classroom quality are notable. In 2006, the average score on the ECERS-R was 4.81 which compared to 3.86 in 2000 (a gain of 1.3 standard deviations). In 2006, almost 90 percent of the classrooms scored above the mean for 2000. The 2006 average score is similar to that found in other studies of publicly funded preschool in this country (Early et. al., 2007). In those areas most likely to be directly related to child learning – Language

and Reasoning, Activities, Interactions, and Program Structure – classrooms on average scored in the good to excellent range.

An average score in 2006 of 3.46 on the SELA also reflects practices that are likely to lead to more learning, with the highest scores in supplying materials that support language and literacy development and in teaching practices that enhance oral language development. Fully, 75 percent scored a 3 or better. However, in the special case of language and literacy, the lower scoring items are mostly related to specific language and literacy skill development, including introducing new vocabulary, assisting children in developing print awareness and letter recognition, supporting phonological development (children's ability to hear the sounds in words) and promoting interest in writing. In addition, assisting parents in supporting their children's language and literacy development and supporting bilingual language development are also lower scoring items.

Results on the PCMI, however, were not so heartening. The only scores above a 3, on a scale of one to five, are on items that reflect the materials in the classroom. Given that the Abbott classrooms are well funded, even these scores seem low and likely represent the same lack of understanding of mathematical learning and teaching shown by the very low scores for teaching support. Six of the seven items that measure whether the teachers actively plan for and support mathematical learning had scores between 1 and 2. Thirty to 50 percent of the classrooms scored a 1 on these items, meaning that none of the desired teaching practices was observed for on these items. Clearly, math learning is enhanced when math is incorporated throughout the classroom activities (Arnold et al., 2002). However, a great deal of math reasoning is also constructed by the

child while using math-related materials (Ginsburg, Inoue, & Seo, 1999). Thus, the slightly better scores on mathematics materials are meaningful, but overall questions remain about whether the program provides enough support for children's learning in this domain to expect large gains.

For some of the low-scoring items on SELA and PCMI that measure fairly specific teaching strategies, it is difficult to judge how much teachers should be expected to use these techniques regularly during the 3.5-4.0 hour observation period. However, over the four years that these measures have been used in Abbott classrooms, there have always been a small percentage of classrooms that score above a 4 on the items, and all classrooms have improved over time. This indicates that it is possible to meet the criteria to score well on these items, and suggests that professional development should continue to focus on these areas.

Of particular interest in these findings is the fact that public school and private child care center classrooms scored the same across almost all measures of quality teaching practices in 2006. In 2000, the private centers scored much more poorly across most measures of quality. This is evidence that standards, funding, and professional development are the conduits to quality and that specific auspices are irrelevant within a public system. Although there is clearly room for improvement, the observational measures indicate that the Abbott Preschool Program is providing good to high quality education to most children. Additional detail is provided by Frede and colleagues (2007).

Effects on Children's Learning and Development

Beginning in the fall of 2005, NIEER and their ELIC partners designed and implemented a two-step research process to estimate the long-term effects of attendance

in an Abbott preschool classroom. The first step was to employ an RDD approach to estimate the effects of the program on children's abilities at kindergarten entry. This approach takes advantage of each district's strict enrollment policy that determines enrollment by the child's date of birth to define the groups. By relying on this assignment rule, one that is unlikely to be related to child and family characteristics, the RDD seeks to reduce the likelihood of selection bias. Thus, rather than compare children who attended and did not attend the program (raising concerns that the family factors that led to this difference might also contribute to differences in learning and development), the RDD approach compares two groups of children who enroll in the Abbott preschool program. One group has had the program and the other is just entering.

One way to interpret the RDD approach is to view it as similar to a randomized trial for children 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. Of course, if only children with birthdays only a few days 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 an assignment variable (age) and measures of children's learning and development. The pre-cutoff sample is used to model 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.

Unfortunately, the RDD approach cannot provide an estimate of effects beyond kindergarten entry. If we employed it a year later, it would provide an estimate of the added effects of kindergarten. Thus, we employ a second approach to obtain estimates beyond kindergarten entry using a conventional no-treatment comparison group, but employ it at kindergarten entry as well so we can compare its estimates to the RDD estimates. If the initial estimates from both analyses are similar, then we have greater confidence in the longitudinal results. If not, then at least we will have an indication of the likely direction and magnitude of the bias in the longitudinal estimates.

Thus, in addition to the two samples drawn for the RDD study we drew an additional comparison sample of kindergarten children who did not attend the Abbott preschool program. We intend to follow both samples of kindergarten children through grade three to assess their performance on measures of academic abilities and the extent to which they are retained in grade or placed in Special Education. As some children attended preschool for one year at age 4 and others attended preschool for two years at ages 3 and 4, we are able to separately estimate the effects of one year and two years of preschool attendance using this second design. In a later follow-up, we will obtain measures of family characteristics, such as mother's education level, language spoken in the home, and family income. However, for the present analyses we the only family background characteristic for which there is a measure is ethnicity. This is less of a

problem than it might be because the communities in our study are fairly homogeneous, all are larger, low-income urban school districts in a single state. In addition, we have ensured that the treatment and comparison samples are balanced with respect to district, and we control for district in the analyses.

Sampling Strategy

In order to limit the logistical costs of data collection, we limited the study to children in the 15 largest Abbott school districts. These districts enroll the vast majority of children in the Abbott preschool program so that distortion introduced is small. However, this strategy likely underestimates the program effect on all children in the state. Previous analyses have shown that classroom quality and children's scores at kindergarten entry are somewhat higher in the 16 smaller districts omitted from the study than in the 15 largest districts. At the beginning of the school year, individual preschool classrooms were randomly sampled from a list of all preschool classrooms in the 15 districts. From each of the randomly sampled classrooms, four children were selected, providing the no-treatment control sample (n= 778) for the RDD. In addition, a kindergarten sample was randomly selected in these districts using a similar procedure. From a list of all kindergarten classrooms, a sample was randomly selected, and then four children who were first time entrants to kindergarten were randomly selected from each classroom. As this was done without consideration of preschool participation, the sample includes a representative proportion of children who did and did not attend Abbott preschool. From this sample of new kindergarten entrants, we constructed the treatment group (n= 766) for the RDD study and the one-year (n= 461), two-year (n= 305), and no-treatment comparison (n= 305) groups for the longitudinal study. The longitudinal study

no-treatment comparison includes a small number of children (n= 59) who attended some non-Abbott preschool program. Overall the sample was 49% female and the ethnic breakdown was 51% Hispanic, 40% African-American, and 8% other. There were no statistically significant differences between the treatment and comparison groups on either gender or ethnicity.

Measurement

Trained research staff from NIEER, William Paterson University, and The College of New Jersey visited each sampled program site, selected children into the sample using a procedure to ensure randomness, and conducted the child assessments as early as possible in the school year. A liaison at each kindergarten site gathered information on the children's preschool status, usually from existing school records but occasionally from parent report, and was reimbursed \$5.00 per sample child. Measures administered to preschoolers and kindergartners in the first year of the study were identical. The battery of child assessments took an average of approximately 25 minutes per child and took place at the child's school program, in a room or quiet area appropriate for assessment.

Receptive Vocabulary. Children's receptive vocabulary was measured using the Peabody Picture Vocabulary Test, 3rd Edition (PPVT-III; Dunn & Dunn, 1997) and for Spanish-speakers the *Test de Vocabulario en Imagenes Peabody* (TVIP; Dunn, Padilla, Lugo, & Dunn, 1986). The PPVT is predictive of general cognitive abilities and is a direct measure of vocabulary size. The rank order of item difficulties is highly correlated with the frequency with which words are used in spoken and written language. The test is adaptive (to avoid floor and ceiling problems), establishing 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. Reliability is good as judged by either split-half reliabilities or test-retest reliabilities. The TVIP is appropriate for measuring growth in Spanish vocabulary for bilingual students and for monolingual Spanish speakers.

All children in our sample were administered the PPVT, regardless of home language, to get some sense of their receptive vocabulary ability in English. All children who spoke some Spanish were also subsequently administered the TVIP. The testing session was then continued, with the additional measures administered in either English or Spanish, depending upon what the child's teacher designated as his or her best testing language.

Mathematical Skills. Children's early mathematical skills were measured with the Woodcock-Johnson Tests of Achievement, 3rd Edition (Woodcock, McGrew, & Mather, 2001) Subtest 10 Applied Problems. For Spanish-speakers the *Bateria Woodcock-Munoz Pruebas de Aprovechamiento – Revisado* (Woodcock & Munoz, 1990) *Prueba 25 Problemas Aplicados* was used. Subtests of the Woodcock-Johnson are reported to have good reliability.

Print Awareness. Children's print awareness was measured using the Print Awareness subtest of the Preschool Comprehensive Test of Phonological and Print Processing (Pre-CTOPPP; Lonigan, Wagner, Torgeson, & Rashotte, 2002). 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-class and low-income samples

and includes a Spanish version. As the Pre-CTOPP was developed recently, relatively little technical information is available about its performance and psychometric properties. Print Awareness 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 the 36 total subtest items is reported and analyzed.

The skills and knowledge measured by the Pre-CTOPPP (which are predictive of later literacy ability) are expected to be present by the end of the preschool year. Thus, it is not an appropriate test for the end of kindergarten and results are not reported past kindergarten entry. In later years, additional literacy assessments will be administered.

Results

Regression Discontinuity Design. To estimate program effects on children's test scores we conducted a series of RDD analyses to guard against model misspecification. 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, or Other), and age. We allowed slope to differ pre- and post-treatment. 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 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.

The regression discontinuity design relies on a number of assumptions that can be tested. One is that programs must adhere to a fairly strict use of a birth-date cut-off date for program enrollment. Each school district employed a birth-date cut-off date for program enrollment, which varied by district from September 30 through December 31. Fortunately, departures from the selection rule were extremely rare. Thus, we conducted “sharp” regression-discontinuity models that dropped the handful of children in our sample whose birth date information appears to be inconsistent with the birth-date cut-off requirement for their programs.

Another key assumption is that there is that the unmeasured population characteristics do not change with birth date. We cannot directly test this assumption. However, we can assess the extent to which results might be changed as we move away from the birth date cutoff. Thus, we repeated our analyses on two subgroups, 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 subgroups produced highly similar estimates of program effects.

The RDD approach also requires that we correctly modeled functional form. 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). We began analyzing third-

order (cubic) polynomial regression models and found the coefficients for the cubic term and its interaction with the cut-off dummy variable 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 and quadratic interaction terms were not significant. Thus, we dropped the quadratic term and its interaction term for the analyses.

Finally, when interpreting the RDD results it is important to note that 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. In the present study, the response functions appear to be linear and reasonably parallel. Moreover, as birth date cutoffs varied by district, the cutoff in our data set occurs over a range with interpolated points and not a single point. Thus, the estimated effects generalize over a broader range of children than would be the case with a single cutoff even if the assumptions of parallel and linear functions did not hold.

The estimated effects of the Abbott preschool program were significant for all three measures. Attending the Abbott program at age 4 is estimated to increase PPVT scores by 4.57 raw score points. This represents an improvement of about 28 percent of the standard deviation for the control (*No Preschool*) group ($es = .28$). Using the control group to calculate expected growth in PPVT scores over 12 months, we can interpret the estimated program effect as 35 percent more growth over the year in children's average vocabulary scores. The estimated effect on children's early math skills as measured by

the Woodcock-Johnson-III Applied Problems subtest scores was 1.36 raw score points ($es = .36$) and equal to a 41 percent increase in growth over the year. The estimated effect on Print Awareness is 14 percent more items answered correctly ($es = .56$) and equal to 96 percent more growth over the year in print awareness.

Longitudinal Study. Regression analyses estimating program effects were conducted on the longitudinal sample with independent variables for student ethnicity, gender, age, and school district, as well as dummy variables indicating one or two years attendance in an Abbott preschool program. Analyses were conducted on raw scores and standard scores (which are more easily interpreted) for the PPVT, raw scores for the Applied Problems test, and percent correct for Print Awareness. Again, the Stata program was employed, and intra-cluster correlation is taken into account through the estimation of cluster-robust standard errors. The estimated effects at kindergarten entry and the end of kindergarten are reported below in the text and in tables that report scores for each group: no preschool, one year of preschool at age 4, and two years of preschool at ages 3 and 4.

The estimated effects of Abbott preschool on the PPVT were statistically significant at both kindergarten entry and exit. Figure 2 portrays estimated gains in receptive vocabulary at the beginning of kindergarten. One year of the Abbott preschool program at age 4 was estimated to increase PPVT scores by 3.82 standard score points at entry ($es = .21$). Two years of Abbott preschool at ages 3 and 4 was estimated to increase PPVT scores by 7.41 standardized score points ($es = .42$). Figure 3 portrays gains on the PPVT at kindergarten exit when the estimated effect of was 3.39 standard score points ($es = .22$) for one year and 6.24 standard score points ($es = .41$) for two years.

Figure 2. Longitudinal Study Receptive Vocabulary at Kindergarten Entry by Years of Attendance (N=1,038)

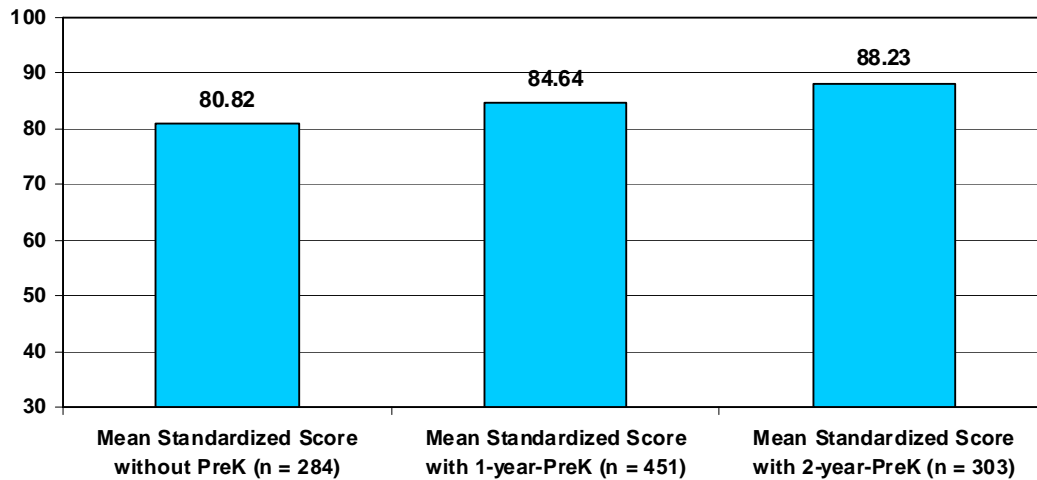
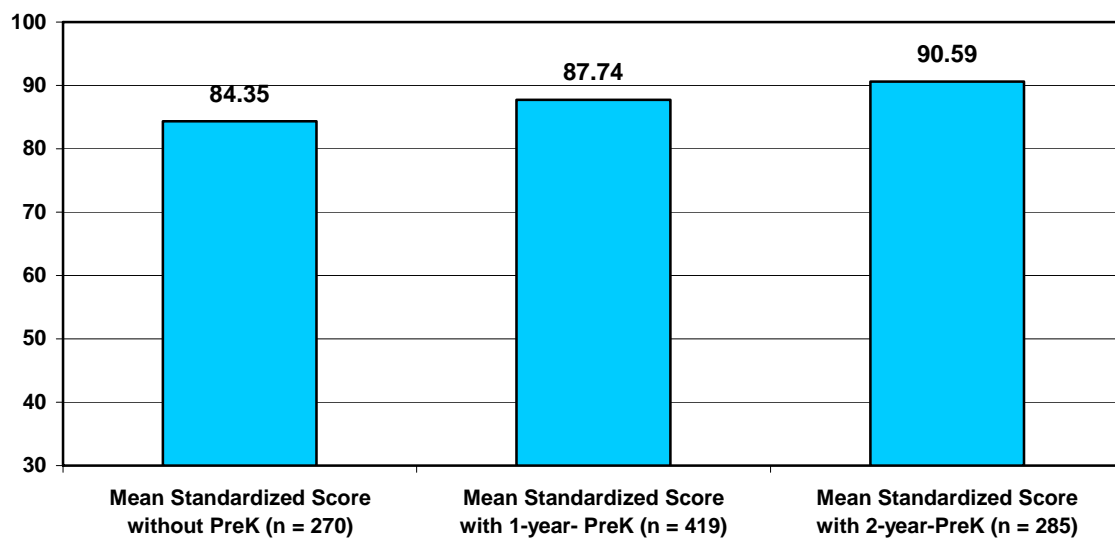


Figure 3. Longitudinal Sample Receptive Vocabulary at the End of Kindergarten by Years of Attendance (N=974)



Estimated effects on children's early math skills as measured by the Woodcock-Johnson-III Applied problems subtest are statistically significant at the start and end of kindergarten. Results are pictured in Figures 4 and 5. The estimated increase in raw score is .86 (es = .20) from one year and 1.47 (es = .34) from two years at the start of

kindergarten. At the end of kindergarten, the estimated effects are .61 raw score points ($es = .13$) from one year and 1.38 raw score points ($es = .29$) from two years.

Figure 4 Longitudinal Study Mathematics Scores at Kindergarten Entry by Years of Attendance ($N=1,054$)

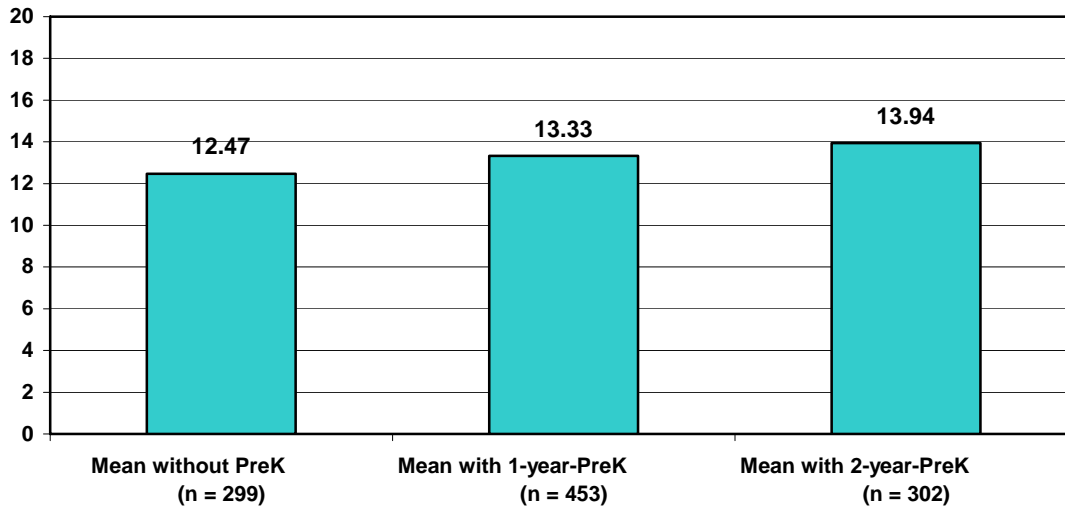
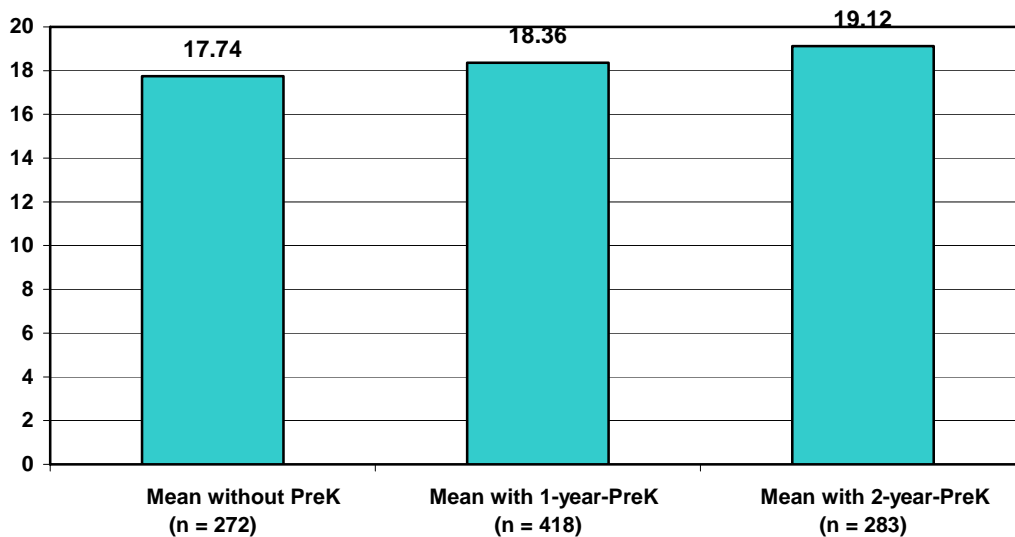


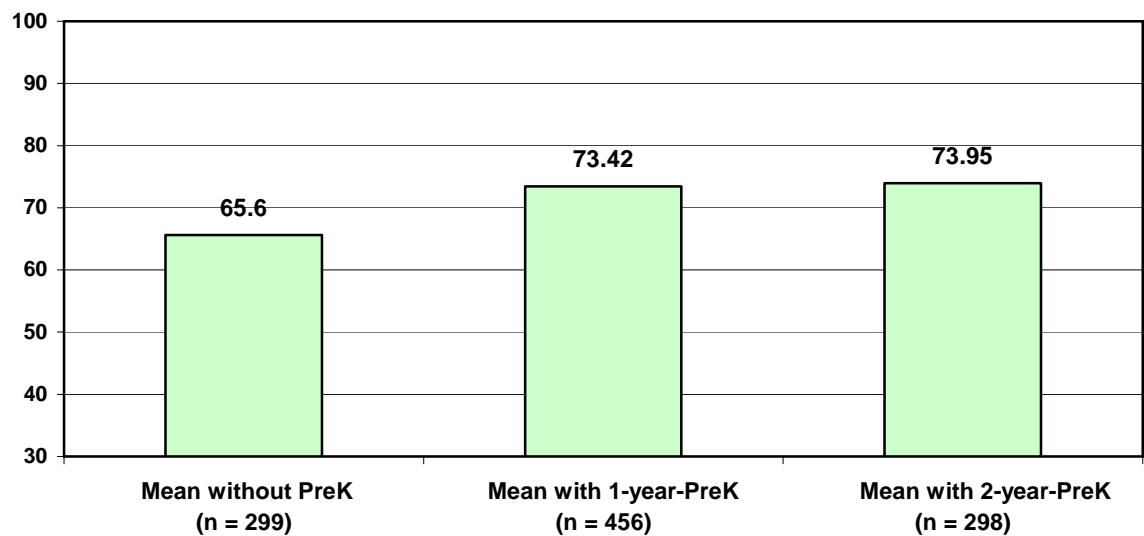
Figure 5. Longitudinal Study Mathematics Scores at the End of Kindergarten by Years of Attendance ($N=973$)



The Print Awareness measure is no longer appropriate at the end of kindergarten, but we estimated effects to see what we might find there as well as at kindergarten entry. At kindergarten entry (Figure 6), the estimated effects of Abbott preschool education

were statistically significant and equaled 7.8 percent more items correct ($es = .29$) for one year and 8.4 percent more items correct ($es = .31$) for two years. By the end of kindergarten, only the estimated effect of two years (1.9%, $es = .14$) remained significant, but it is apparent that many children topped out on the measure. Even the no-treatment group scored on average 91% correct by end of kindergarten, compared to about 93% for the one and two year groups. Although one can interpret these results as reflecting a measurement problem, they can also be viewed as evidence that most children master this skill by the end of kindergarten regardless of previous experiences.

Figure 6. Longitudinal Study Print Awareness Scores at the Kindergarten Entry by Years of Attendance (% correct) (N=1,053)



In order to compare the RDD and longitudinal results, we re-estimated the effects of Abbott preschool education on the PPVT at kindergarten entry using raw scores. In the raw score analyses, one year 4 was estimated to increase PPVT scores by 4.06 raw score points. This effect is slightly smaller (11 percent) than the estimated 4.57 point gain found using the RDD approach. However, using the longitudinal study comparison group the immediate estimated effects of one year were reduced by 37% for math (.86

versus 1.36 RDD) and 44% for print awareness (7.8% versus 14% RDD). This suggests that sample selection bias may be leading to substantial underestimation of program effects in the longitudinal study.

Discussion

Considerable attention and resources have been invested in the Abbott Preschool Program. The Abbott program's relatively high per pupil cost of around \$11,000 reflects not only New Jersey's high cost of living (K-12 costs \$14,000 per child), but high program standards. The program operates for a full school day, employs licensed teachers paid on the same scale as public school teachers, has a maximum class size of 15 with an assistant teacher assigned to each classroom, and has dedicated staff to work with parents (Barnett, Hustedt, Hawkinson and Robin, 2006). It is also notable that most classrooms are operated by private providers under contract to the public schools. The population served by the Abbott districts is primarily low-income families of Hispanic and African-American origins. Thus, there is a great deal of interest in how effective Abbott classrooms are in helping to improve the knowledge, skills and dispositions of these children as they enter kindergarten and proceed through school. This paper brings together evidence from several different types of analyses to address this question.

The effectiveness of any program depends on how well it is implemented, but quality of implementation is not always in program evaluations. In this study, we assessed program quality using multiple instruments. These assessments indicated that program quality was generally good and that this was considerably higher for the cohort of children in our study than it was several years earlier before full implementation of Abbott preschool standards. The quality assessments were not so high as to preclude

further improvements. The level of quality observed led us to expect moderate gains for children. Somewhat smaller gains might be expected in math given the relatively poor performance of teachers on the math observation. However, it may also be the case that the children had fewer supports for math learning (compared to language and literacy) outside the classroom. The scores on the most commonly used assessment of classroom quality are similar to those found for many other public preschool education programs, which supports some generalization of the results from this study.

Earlier studies had indicated that the Abbott Preschool Program has beneficial effects on children's skills at kindergarten entry, including a study employing RDD on a somewhat broader sample of Abbott districts (Wong, Cook, Barnett, Jung, & Lamy in press; Frede et al, 2004; Lamy et al., 2005). We find positive effects on children's learning in the areas of oral language, literacy and math skills with effect sizes that range from .20 to .56 depending on the measure and research design. The effects of two years are found to be significantly and substantially larger for language and math, but not for print awareness. Despite evidence of downward bias in the longitudinal study estimates, we found that substantial gains in language and math persisted through the end of kindergarten. By the end of kindergarten, the test of print awareness was not really appropriate anymore, and most children had mastered the print awareness skills tested. Thus, conclusions regarding the persistence of program effects on literacy other than oral language cannot really be drawn.

Children's early print awareness and receptive vocabulary skills have been found to predict later reading abilities in the early elementary grades (Snow, Burns, & Griffin, 1998). In addition, the effects found in this study are the first link in a chain that can

produce the long-term school success and economic benefits found in other preschool education studies that have followed children into adulthood (Schweinhart et al., 2005; Campbell et al., 2002; Reynolds, Temple, Robertson, & Mann, 2002).

We conducted two separate studies of program effects at kindergarten entry to address concerns that the simple comparison of children who attended and did not attend Abbott preschool programs might be biased by unmeasured differences between the groups. This does seem to have been the case. The regression discontinuity design which attempts to control for these unmeasured differences provides estimates for the effects of one year of preschool education that are higher by 11 percent for language (PPVT), 37 percent for math, and 44 percent for print awareness. This indicates that the estimated effects in our longitudinal study underestimate the effects of preschool by meaningful amounts.

Thus, the longitudinal study addresses the question of the extent to which effects may fade out over time, but it must be understood that the longitudinal study somewhat underestimates the effects of the Abbott preschool program. At least for PPVT, the underestimation appears to be fairly modest. Results of the study indicate that there are persistent effects on children's learning through the end of kindergarten, with only modest declines in the advantages from attending Abbott preschool programs for language and math. (Curiously, the largest decline was for the effect of one year of preschool education on math scores and the smallest decline for the effect of two years of preschool education on math scores.) Tests at the end of first grade will include measures of literacy more broadly.

Very little research exists that compares the effects of one year versus two years of preschool attendance. Children who attended the Abbott Preschool Program for two years at ages 3 and 4 out-perform children who attended for only one year at age 4 and those who did not attend on all of the outcome measures with one exception. The gains in language and math from two years are quite large, nearly double for language and 70 percent larger for math. Children who had two years of preschool do not score significantly differently from those who had one year on the Print Awareness test. This is not a great surprise since this test is actually designed to assess preschool children's preschool literacy skills, and the majority of the children score well on it by the end of kindergarten. Caution must be used in interpreting these results. We cannot control for possible selection bias across the groups. Parents who know about and choose to send their children to preschool at age 3 may be different in immeasurable ways from those who only send them at 4. For this comparison we do not have the estimates from the more rigorous RDD to verify our results. The fact that this study is large scale and that it is fairly safe to assume that the quality of program for both years is similar adds to the importance of the findings.

Future Research Directions

We plan to continue the present study through the end of third grade. In these future studies, it will be important to add literacy skills measures that are age appropriate and to expand the data collection to include information on grade retention and special education placement as well as achievement test results. In addition, it would be useful to add a measure of social and emotional development, as recent studies have shown that well-designed preschool education programs can produce positive gains in self-regulation

and other areas of development that are at least as important for later life success as the domains measured by achievement tests (Diamond, Barnett, Thomas, & Munro, 2007).

One aspect of the Abbott Preschool program that is unusual is that it serves entire communities with high percentages of children from low-income families rather than targeting individuals who are from low-income families. This approach may have resulted in differences in who participates, and not just in permitting the participation of children from higher-income families. Also, the inclusion of children from a broader range of backgrounds could have improved the preschool learning experiences for children from low-income families, while the participation of the vast majority of children from the community in the program could have improved the kindergarten learning experiences (for example, improving overall classroom climate or allowing teachers to spend less time on remediation or even raise the overall level of classroom interactions). Future studies that could shed light on these questions would undoubtedly be useful.

New studies that focus on the contributions of quality and quantity to program effectiveness would seem to be desirable. This program appears to have larger effects than found in many studies for typical early education and child care programs, but smaller effects than found for some others. In this current study we did not have sufficient resources to tie the classroom quality scores in preschool to the child outcome data. Future research designed to do this would help in determining the relationship between level of implementation and program effectiveness. Randomized trials that studied both program standards and observed program quality (perhaps using coaching or other approaches to professional development to induce changes in quality) would be

especially useful. A better understanding of the dosage of specific math and literacy classroom practices that is necessary to produce large gains in children's learning might be obtained from such studies. Prospective studies comparing the effects of one year of preschool education at age four versus two years of preschool education starting at age three also would appear to be warranted.

Our study suggests that selection bias can be a problem in the commonly used nonequivalent comparison group studies. However, this is hardly a strong result or one that is readily generalized given the lack of variables that might have served as good statistical controls for family background or a pre-test measure of children's abilities. Of course, it is not uncommon for preschool program evaluations to lack pre-test measures because of the difficulties of identifying children who do not attend a preschool education program prior to kindergarten entry. Additional studies comparing results of randomized trials, RDD, and more typical quasi-experimental designs using overlapping data for the same program and population across would be extremely useful.

Implications for Policy

The results of this study add to the considerable body of evidence indicating that quality preschool education can make significant contributions to efforts to improve children's learning and development (Frede, 1998). It also adds to the evidence that substantial benefits persist at least through the end of kindergarten. Moreover, this study confirms that such effects can be produced with today's children on a large scale by a public program administered through the public schools, reinforcing recent findings from Tulsa, Oklahoma (Gormley, Gayer, Phillips, & Dawson, 2005). The population in this study was largely minority, heavily Hispanic, and mostly low-income. However, the

program was not targeted to individuals, but to communities with large percentages of children from low-income families. This may or may not have contributed to its success, but it appears to be one effective strategy. The Abbott Preschool Program had high program standards, and the contrast between its estimated impacts and those of more typical child care and other programs should give policy makers one more reason to be circumspect about the potential for programs with lower standards to produce similarly strong results (Magnuson, Ruhm, & Waldfogel, 2007; NICHD Early Child Care Research Network, 2006).

This study establishes that public programs can produce reasonably strong results using private providers. States and localities that find it difficult to expand the public school system to provide preschool education due to space limitations and other constraints should consider a mixed delivery system using public school, private schools and child care centers, and Head Start agencies. Private programs can provide equally effective preschool education so long as they are held to the same high standards as public schools and receive public funds adequate to meet those standards. In the Abbott program, there is oversight from the state and local schools that provides financial and educational accountability (Frede, 2005). Whether other approaches that grant private programs more autonomy and depend on parent choice in the market for educational accountability would produce similar results is unknown.

Two of the studies most widely cited in support of public investments in preschool education are the Perry Preschool and Chicago Child Parent Center studies (Schweinhart et al., 2005; Reynolds et al, 2007). In both of these studies, the preschool programs provided most children with two years of education beginning at age three.

Our study suggests that it would be unwise to expect similar results from one-year of preschool education, even if other characteristics of the program were equivalent. Policy makers who are seeking to enhance the effectiveness of their investments in preschool education, including those focused on decreasing the achievement gap between advantaged and disadvantaged children should consider serving children for at least two years beginning at age three. Other studies suggest that this is not the only quantity issue relating to effectiveness, and that length of the day and school year are also deserving of scrutiny by policy makers seeking better educational outcomes (Robin, Frede, & Barnett, 2006).

One final lesson for policy-makers from this study is that it can be important to avoid rushing to evaluate the effectiveness of a new program. Programs take time to develop and to be implemented as intended. Initial program evaluations might best focus on determining whether the program is implemented as designed and how well it is actually delivering education. Such studies provide valuable information to those responsible for developing the program. Once an acceptable quality level is attained, including quality of classroom practice, then evaluation of the effects on children's learning and development can reveal the extent to which implementing the program as planned achieves desired goals. Premature evaluation of outcomes without attention to implementation could easily find only that an inadequately developed, partially or even poorly implemented program was not very effective.

References

- Arnold, D.H., Fischer, P.H., Doctoroff, G.L., & Dobbs, J. (2002). Accelerating math development in Head Start classrooms. *Journal of Educational Psychology, 94*, 762-770.
- Barnett, W.S., Hustedt, J.T., Hawkinson, L.E., & Robin, K.B. (2006). *The state of preschool 2006: State preschool yearbook*. New Brunswick, NJ: National Institute for Early Education Research.
- Barnett, W. S., Jung, K., Lamy, C., Wong, V., Cook, T. (2007, March). *Effects of five state prekindergarten programs on early learning*. Paper presented at the bi-annual Society for Research in Child Development, Boston, MA.
- Barnett, W.S., Lamy, C., & Jung, K. (2005). *The Effects of State Prekindergarten Programs on Young Children's School Readiness in Five States*. NIEER Policy Report. New Brunswick, NJ: National Institute for Early Education Research.
- Barnett, W. S., Tarr, J., Esposito Lamy, C., & Frede, E. (2002). *Fragile Lives, Shattered Dreams: A Report on Implementation of Preschool Education in New Jersey's Abbott Districts*. Rutgers University, New Brunswick, NJ: CEER.
- Burchinal, M.R., Cryer, D., Clifford, R.M, & Howes, C. (2002). Caregiver training and classroom quality in child care centers. *Applied Developmental Science, 6*, 2-11.
- 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., & Campbell, D. T. (1979). *Quasi experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin.

- Diamond, A., Barnett, W.S., Thomas, J., & Munro, S. (2007). Preschool program improves cognitive control. *Science*, 318, 1387-1388.
- Dunn, L. M. & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test-Third Edition (PPVT-3)*. Circle Pines, MN: AGS Publishing.
- Dunn, Padilla, Lugo & Dunn, (1986). *Test de Vocabulario en Imágenes Peabody (TVIP)*. Circle Pines, MN: AGS Publishing.
- Early, D.M. et al. (2007). Teachers' Education, Classroom Quality, and Young Children's Academic Skills: Results From Seven Studies of Preschool Programs. *Child Development*, 78 (2), 558-580.
- Frede, E. (1998) A sociocultural analysis of the long-term benefits of preschool for children in poverty. In Barnett, W.S. and Boocock, SS (Eds) *Early Care and Education: Lasting Effects for Children in Poverty*. Buffalo, NY: SUNY Press.
- Frede, E. (2005) Assessment in a continuous improvement cycle: New Jersey's Abbott preschool program, invited paper for the National Early Childhood Accountability Task Force with support from the Pew Charitable Trusts, the Foundation for Child Development and the Joyce Foundation.
- Frede, E., Lamy, C.E., & Boyd, J.S. (2005) Not Just Calendars and Counting Blocks: Using the NAEYC/NCTM Joint Position Statement "Early Childhood Mathematics: Promoting Good Beginnings" as a Basis for Measuring Classroom Teaching Practices and Their Relationship to Child Outcome a paper presented at the annual National Association for the Education of Young Children conference, Washington, DC.

- Frede, E., Lamy, C.E. with Seplocha, H., Strasser, J., Jambunathan, S., Juncker, J., & Wolock, E. (2004). *A rising tide: Classroom quality and language skills in the Abbott Preschool Program: Year Two Preliminary Update of the Early Learning Improvement Consortium*. Trenton, NJ: New Jersey Department of Education. www.nj.gov/njded/ece.
- Frede, E., Weber, M., Hornbeck, A., Stevenson-Boyd, J., & Colon, A. (2005). *Preschool Classroom Mathematics Inventory*. Available from the first author at efrede@nieer.org.
- Ginsburg, H.P., Inoue, N., & Seo, K.H. (1999). Young children doing mathematics: Observations of everyday activities. In Copley, J. (Ed.) *Mathematics in the Early Years*. Washington, D.C: NAEYC.
- Gormley, W.T., Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal pre-k on cognitive development. *Developmental Psychology*, 41(6), 872-884.
- Harms, T., Clifford, R., & Cryer, D. (2005). *Early Childhood Environment Rating Scale (ECERS-R)*, revised edition. New York, NY: Teacher College Press.
- Lamy, C., Frede, E., & ELIC. (2005). *Giant Steps for the Littlest Children: Progress in the Sixth Year of the Abbott Preschool Program*. Trenton, NJ: New Jersey Department of Education. www.nj.gov/njded/ece
- Lonigan, C., Wagner, R., Torgeson, J. & Rashotte, C. (2002). Preschool Comprehensive Test of Phonological & Print Processing (Pre-CTOPPP). Tallahassee, FL: Florida State University, Department of Psychology.
- NAEYC & NCTM. (2002). *Early childhood mathematics: Promoting good beginnings*. A joint position statement of the National Association for the Education of Young

- Children (NAEYC) and the National Council for Teachers of Mathematics (NCTM). Available at: <http://www.naeyc.org/about.positions/psmath.asp> or <http://www.nctm.org/about/content.aspx?id=6352>.
- NJ DOE (2002b). *New Jersey Preschool Teaching and Learning Expectations: Standards of Quality*. Trenton: author.
- NJ DOE (2002a). *New Jersey Kindergarten Core Curriculum Content Standards*. Trenton: author.
- Reynolds, A.J., Temple, J.A., Ou, S., Robertson, D.L., Mersky, J.P., Topitzes, J.W., & Niles, M.D. (2007). Effects of a school-based, early childhood intervention on adult health and well-being: A 19 year follow-up of low-income families. *Archives of Pediatrics and Adolescent Medicine*, 161(8), 730-739.
- Robin, K., Frede, E. & Barnett, W.S. (2006) Is more better?: The effects of full-day vs. half-day preschool on early school achievement. <http://nieer.org/docs/index.php?DocID=144>
- 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.
- Smith, S.; Davidson, S & Weisenfeld, G (2001). *Supports for Early Literacy Assessment for Early Childhood Programs Serving Preschool-Age Children*. New York: New York University
- Snow, C., Burns, M. S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.

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.

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.