Improving Early Education Programs through Data-based Decision Making

By Shannon Riley-Ayers, Ellen Frede,
W. Steven Barnett and Kimberly Brenneman

NIEER
NATIONAL INSTITUTE FOR EARLY EDUCATION RESEARCH

RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

Contents

Executive Summary

As publicly funded preschool education grows, states are moving toward establishing accountability systems that better measure program effectiveness and therefore how effectively the public's money is spent. When well-conceived, these systems not only serve as report cards on programs to policymakers and the public, they also provide data important for continuous program improvement. Reliable guidance on how states can best study program effectiveness is limited. In fact, majority of state evaluations of preschool programs are less than rigorous in terms of scientific standards with many having such flaws that there are severe limitations in interpreting their results (Gilliam & Zigler, 2000, 2004).

Challenges are presented by the fact that children are in the early stages of development, programs vary widely, and accountability goals and therefore research considerations may vary from state to state. Thus the approaches states take in studying programs and their effectiveness require careful consideration and this paper will shed light on state's options for this important and necessary evaluation.

In this report we present five options for studying program effectiveness. They are summarized in chart form at the end of this executive summary. The summary chart provides a quick glance at the research questions that each design addresses, a brief outline of the methodology, the main issues or concerns with the design, and the positive aspects of the design approach. An estimate of costs for each evaluation can be found in the full document that follows this executive summary. At the end of the chart we describe additional research questions that examine quality and can be added to any of

the research designs excluding the first: Utilizing Extant Data. We also present an economic analysis study as the final piece of the chart.

1. Utilizing Extant Data

This study approach uses extant data to answer the research questions. It taps into the differences at kindergarten entry and third grade in children that attended the preschool program and those who did not attend. However, as can been seen in the chart, there are several issues with this approach. The most striking issue is the appropriate determination of children who fit in the two groups (preschool treatment and no preschool), keeping in mind the likelihood that these two groups may or may not differ inherently before attending preschool.

2. Nonequivalent Groups, Post-test Only (Kindergarten)

The second design outlined in the chart is the non-equivalent groups, post-test only. This study design answers specific questions about children's academic achievement and social skills at kindergarten entry and over time. This study approach examines the children beginning at kindergarten entry and, like the extant data approach, creates the two groups — those children who attended the preschool program and those who did not. As with the extant data approach, this can create issues of selection bias where the groups may differ inherently. Nevertheless, this approach provides a good look at the same students over time and does so with specific standardized child assessments that provide a report of academic achievement and social skills.

3. Nonequivalent Groups Pre- and Post-test (Preschool)

The third design, nonequivalent groups with a pre- and post-test, is similar to the above design with the exception that the children are selected and assessed before the beginning of the preschool program. This approach requires either a waiting list for entry into the program or some other type of screening measure to determine eligibility so that the two groups (one receiving the treatment of preschool and one not receiving it) can be determined and both groups assessed at this time. This approach enables investigators some reduction in selection bias issues and provides more statistical power for reliable identification of modest effects. It requires one additional year of assessments than the post-test only option. This extra year of data collection creates a higher end cost because of four years of collecting data versus three years.

The three options described above and detailed in the table rely on identifying non-equivalent groups for comparison. They are nonequivalent because some attended preschool and some did not, but this did not happen randomly. Selection bias becomes an issue when using groups that self-select to participate in the preschool program and/or who are selected based on program eligibility criteria. Selection bias in this case relates to the concern that the two groups could be inherently different *in ways that are not measured and are related to children's learning and development* at the start of the study before the treatment (the preschool program) is provided. If such bias exists, the estimates of the effects of the program are likely biased. Selection bias could for example mask systematic differences between the educational aspirations of parents of the two groups. This would undoubtedly lead to the groups performing differently in

school. In studying preschool programs that serve disadvantaged children, selection bias most often appears to underestimate program effects.

4. Regression-Discontinuity Design

The strongest research design next to randomized trials is one that offers protection against selection bias beyond simply controlling for family background in a statistical analysis. This is the regression-discontinuity design (RDD). It employs a statistical model that uses stringent age cut-offs to define groups. Testing groups using the age cut-offs and then statistically adjusting for age variation reduces the likelihood that selection bias has an appreciable impact on study results. For states considering this approach, it is important to have a sufficient number of children enrolled to provide a large sample to provide confidence in the estimates from an RDD study. For this approach to be most successful, it is best conducted when there is no significant expansion in the provision of the preschool program from the prior year in the communities participating in the study. Confidence in RDD studies is also increased if children can be assessed very early in the school year.

RDD studies can be combined with an ordinary nonequivalent comparison (NC) group longitudinal study (2 above) of children who did and did not attend preschool. In this case, the children are followed from kindergarten to third grade. The combination of these two approaches enables investigators to assess the extent to which selection bias affects the estimates at kindergarten entry. If sufficient family background and school attendance can be obtained, it may be possible to model some of the differences that contribute to the bias and thereby reduce that bias. The NC longitudinal design adds to the study the ability to estimate effects on social and emotional development at the end of

kindergarten, and to estimate effects on achievement, grade retention, special education, and socio-emotional development into the elementary years. These outcomes are important for increasing confidence in estimates of the economic value of the program's benefits.

5. Randomized Trial

The best approach to preventing selection bias is to employ a randomized trial design. Children are randomly assigned to either attend preschool or not. One complication of this approach is that it would require a lottery system or some other means of determining who is enrolled and who is not. This is not always possible. Two circumstances that work well for randomization is when there is a waiting list for preschool enrollment or where there is the potential for an expansion of services to serve a group not currently eligible for the program.

Examination of Quality

States should consider conducting preschool evaluations in addition to studies of program effects. This requires examining the preschool classrooms with established measures in order to look at the quality of the classrooms. This provides a clear report of the status of the classrooms across the program and can be conducted by a representative sampling of classrooms rather than examining all the classrooms in the program. Analyses can determine the number of classrooms that would need to be observed and evaluated in order to provide a reasonably representative reflection of the classrooms across the program.

Economic Analysis

It is also advisable for states to consider cost studies in conjunction with some type of broader economic analysis. The level of intensity of the cost analysis should depend on the state's purposes and the level of precision needed. A simple paper and pencil or web-based survey of providers together with existing state data might provide sufficient information for a rough cost analysis. A decision about an analysis of benefits could wait until after estimates of program effects are available. Based on the results a choice could be made between simple extrapolations and more complex analyses that directly build upon detailed effectiveness results.

DESIGN SUMMARY CHART

Research Design	Research Question(s)	Brief Outline of Method	Main Issues/Concerns	Positive Aspects
1. Utilizing Extant Data	(1) How do children who attended preschool compare on a kindergarten measure to children who did not attend the program? (2) How do children who attended preschool compare on standardized testing in third grade to children who did not attend the program?	• Compile existing data • Create two groups: one of children who did attend the preschool program and one of children who did not attend the program • Match the groups on family demographics • Conduct statistical analyses	• Tends to underestimate the effects of preschool programs that target disadvantaged populations • Collection of data through paper records • Inaccuracy of preschool attendance records or reports • Selection bias- the group of children who attended preschool are inherently different than the control group • Matching groups • Provides information about the preschool program in existence 4 years earlier when looking at the third grade data	• No additional assessments are conducted • Timely results available

Research Design	Research Question(s)	Brief Outline of Method	Main Issues/Concerns	Positive Aspects
2. Nonequivalent Groups, Post-test Only (Kindergarten)	(1) What effect does preschool have on children's academic achievement and social skills at kindergarten entry? (2) How do children who attend preschool compare over time in academic achievement and social skills with children who do not attend the program? (3) What effect does preschool have on grade retention and special education status?	• Compares children in the same age cohort starting in kindergarten who did and did not attend the preschool program • Child assessments at kindergarten entry and the end of kindergarten, first, second and third grades • Monitor children's grade retention and special education status	• Selection bias that can lead to underestimation of the effects of the preschool program • Matching groups on family characteristics • Type of preschool experiences of the control group • Child tracking over time/attrition	• Estimate of long-term effects

Research Design	Research Question(s)	Brief Outline of Method	Main Issues/Concerns	Positive Aspects
3. Nonequivalent Groups, Pre- and Post-test (Preschool)	(1) What effect does preschool have on children's academic achievement and social skills at kindergarten entry? (2) How do children who attend preschool compare over time in academic achievement and social skills with children who do not attend the program? (3) What effect does preschool have on grade retention and special education status?	• Identification of children at the beginning of the preschool program • Pre-test children accepted into the preschool program *and* those on waiting lists before the start of the preschool year • Child assessments at the end of preschool, kindergarten entry and the end of kindergarten, first, second and third grades • Monitor children's grade retention and special education status	• Must have a waiting list or some other method to screen applicants to determine eligibility • Matching groups on family characteristics • Type of preschool experiences of the control group • Child tracking over time/attrition	• Pre-test increases the statistical power to allow for reliable identification of modest effects • Estimates of effects immediate and long term

Research Design	Research Question(s)	Brief Outline of Method	Main Issues/Concerns	Positive Aspects
4. Regression-Discontinuity	(1) What effect does preschool have on children's academic achievement at kindergarten entry?	• Compare two groups of children who enroll in the preschool program voluntarily • Compare the kindergarten group as the treatment group and the preschool group as the control group at the start of the year • Sophisticated analyses are conducted on the data	• Cannot examine social skills • Requires greater expertise in statistical methodology	• Provides a control for selection bias • Can be combined with a longitudinal approach to provide a statistical estimate of the impact of selection bias in a nonequivalent comparison group design • Provides immediate results

Research Design	Research Question(s)	Brief Outline of Method	Main Issues/Concerns	Positive Aspects
5. Randomized Trial	(1) What effect does preschool have on children's academic achievement and social skills at kindergarten entry? (2) How do children who attend preschool compare over time in academic achievement and social skills with children who do not attend the program? (3) What effect does preschool have on grade retention and special education status?	• Random assignment to the preschool program • Child assessments pre- and post- preschool (both children who attended the program and those who did not attend), kindergarten entry and the end of kindergarten, first, second and third grades • Monitor children's grade retention and special education status	• Need random assignment to the preschool program • Type of preschool experiences of the control group • Child tracking over time/attrition	• Should eliminate selection bias • Estimates long term effects

Research Design	**Research Question(s)**	**Brief Outline of Method**	**Main Issues/Concerns**	**Positive Aspects**
Examination of quality:	(1) How does the quality of preschool classrooms differ across auspice, provider, or teacher? And/or (2) What is the impact of quality of the preschool experience on student outcomes?	• Observation of preschool classrooms for quality by an outside observer	• Additional cost and time to the evaluation • Potential increase of sample size	• Quality is an important component to summarize when reporting on a preschool program • Provides a look at the improvement of the program over time • Can control for levels of quality in programs therefore increasing precision

Research Design	**Research Question(s)**	**Brief Outline of Method**	**Main Issues/Concerns**	**Positive Aspects**
Economic Analysis	What is the comparison of the costs of the program to the savings that are attributable to the outcomes of said program?	• Begin with a cost analysis to determine a cost estimation • Conduct a cost-benefit analysis • May decide to estimate benefits by relying on existing longitudinal research so that results are produced more quickly than conducting your own longitudinal research	• Can be very difficult, if not impossible, to produce credible measures of monetary value of some outcomes • Most benefits of preschool education for which a dollar value can be estimated occur far into the future and, thus, would take a long time for study results	Can contribute to internal and external decisions such as: • Planning budget allocations and projecting resource needs • Improving the efficiency of program operation • Setting fair and adequate fee or payment schedules • Identifying the impacts of and finding the best ways to meet regulatory and licensing standards • Influencing decisions and judgments made by people external to the program

Introduction

Policy makers, the early childhood profession, and other stakeholders in young children's lives share the responsibility to regularly engage in program evaluation (NAEYC & NAECS/SDE, 2003). Prior to charting a course for program evaluation, however, state officials must consider the purposes of such evaluations and the audience(s) to which they are addressed (Frede, 2005). Purposes for program evaluation may vary from obtaining data that can be taken into account in high stakes decision-making such as determining program funding to measuring programs and/or children's progress in them for reporting and program improvement purposes. Audiences may include policymakers, educators, researchers and the public in general. Whatever the case, well-conceived program evaluation is a valuable source of information with which to inform policy, teaching practice, and a continuous cycle of program improvement.

Both the purpose of program evaluation and the audience(s) affect what will be measured and how it will be measured. If the legislature wants to know whether the money is well spent, then accountability may include expenditure analyses as well as child outcome studies. For that to happen, however, it is critical that the program standards and outcomes desired already be established.

If, on the other hand, the only accountability issue is child outcomes then only child learning standards need be considered. Where the relationship between classroom implementation and child progress is of interest, then curriculum and teaching must be examined. Program standards that detail criteria for program operations such as administrator credentials or community involvement would be necessary if program implementation is being assessed. Standards at the child, site, district, and state level

may be utilized. As accountability and evaluation data is collected states may consider revisiting and raising standards as improvement is seen.

Framing the Central Research Questions

To assist state officials in determining the overall evaluation approach for their state, the Report of the National Early Childhood Accountability Task Force (Schultz & Kagan, 2007) outlines four approaches that respond to different questions of interest. Each varies in its rigor and presents its own set of challenges and cautions. It is based on this information that state officials should begin to formulate the types of research questions they wish to address. Then, based on the general assessment approach that is chosen, a research design can be selected along with the methodology, instrumentation, and data analyses.

One approach is to consider what the status is of all children in the state through the child population approach. To accomplish this, a state would sample from the population a representative group, including children who are not enrolled or did not attend the preschool program. If a large enough sample is selected then information can be summarized about local agencies. Data on children's progress in learning and development is collected and can be done so at the preschool level, at kindergarten entry, and/or tracked annually over time. Two issues that arise with this approach are collecting the sample, especially locating those children who do not attend the preschool program or move several times. Another is the danger that policymakers may inappropriately utilize this child-level information to retrospectively make inferences about the quality of preschool experiences.

The Accountability Task Force describes a program population approach to examine the quality of services in the preschool program. This approach does not use a sampling procedure, but rather uses program population data from all programs. This data includes program quality, workforce, and public investments. This approach can demonstrate if changes have taken place, but cannot necessarily provide the reason for the change.

The local agency quality approach is also described in the task force's report and this considers quality at the local agency. This local data provides state managers with information to use in working with the agencies to improve quality. The intention with this approach is to work with individual agencies on meeting standards and moving to higher levels of quality over time. What is important to manage here is the examination of the programs and providing technical assistance based on the deficiencies in a program.

The approach that is the focus for this paper is the state program evaluation approach. This addresses the quality of the program and how children are progressing. This approach combines child and program data which enables state officials to describe the relationship between the program, practices, quality, and child outcomes. This method utilizes the sampling approach that represents the universe of programs being studied. It examines child outcome data on learning and development, child characteristics, and data on the centers including background data on teachers, staff, classroom practices, and quality. To counter some issues that arise there must be consideration of the funding for the local agencies because many receive money from

several sources, consideration of the amount of treatment (preschool) received (attendance issues), and finding the appropriate control or comparison group.

In addition, longitudinal information should be considered an important component of any program evaluation. The need to know the long-term effect of preschool is an important factor to consider when choosing an approach. This longitudinal look at the effects of preschool requires some continuity in program accountability and improvement from preschool to grade 3 and even beyond. The Report of the National Early Childhood Accountability Task Force (2007) recommends that states link and align preschool to grade 3 accountability efforts by creating a vertically aligned framework of standards for child learning and program/classroom quality and developing a unified system of child identification numbers that would allow tracking of children's demographic characteristics, program experiences, and assessment information across the years. This partnership between early childhood and elementary education provides the foundation for creating a shared responsibility for children's success between the two entities.

Adequacy of Design

Just as states must decide their overall approach to accountability based on the questions they would like to answer, their choice of the study design will determine the rigor of the research conducted and dictate the strength of the results. The designs outlined below are listed in terms of rigor:

(1) When ethical and feasible, a randomized trial is the best method for answering well-defined questions about "what works" (Feuer, Towne, & Shavelson, 2002). A randomized trial uses a lottery system to randomly assign children to the treatment group

(preschool program) or the control group (no attendance in the preschool program). This provides the best way to create equivalent groups and offers the strongest support for cause and effect between the treatment and the outcomes. Any outcome differences that are observed between the groups at the end of the evaluation are likely to be due to treatment and not due to differences that existed between the groups at the start of the study (Shadish, Cook, & Campbell, 2002).

Even small, local randomized trials can provide more accurate information than large non-experimental studies, particularly when results can be compared across multiple small trials with somewhat different programs, populations, and contexts. Such replication is important for understanding how program outcomes depend on what is provided, who is served, and other circumstances (for example, K-12 policies or economic conditions). This approach is often difficult, time-consuming, and expensive.

(2) Next best are studies with quasi-experimental designs using regression discontinuity that are often more feasible and economical. This type of experiment lacks random assignment, but steps are taken to ensure comparability of treatment and control groups. These studies are specifically designed to disentangle family influences from program influences.

(3) Third in this progression are prospective longitudinal studies specifically designed to study natural variation in programs and children's participation Typically, data are directly collected on programs and children, including data on the abilities of children attending and not attending when they begin preschool education — and not just after the program.

(4) Least rigorous but potentially useful in the right circumstances are studies using survey data where preschool program participation is based on retrospective parental report. These studies commonly produce estimates of the effects of programs like Head Start that are contradicted by results from nationally representative randomized trials. In other words, they fail the best available test of their ability to produce accurate estimates of program outcomes. Such studies are poor sources of information about causal questions (e.g., what works).

Adequacy of Assessment

A research study design dictates the strength of the results, but the assessments chosen as part of that study will dictate the content of the results and impact the reliability and validity of the study results. NAEYC and NAECS/SDE (2003) jointly agree that ethical, appropriate, valid and reliable assessments should be made a central part of all early childhood programs. They state that these assessments should be connected to specific and beneficial purposes: "(1) making sound decisions about teaching and learning, (2) identifying significant concerns that may require focused intervention for individual children, and (3) helping programs improve their educational and developmental interventions" (p. 2).

There are several types of assessments available to researchers and each type has a specific purpose. For instance, not all assessments lend themselves to large-scale research studies and surely not all assessments are appropriate measures to give to all children in a classroom to inform instruction. To begin, though, all instruments should demonstrate reliability and validity for their intended use. Issues arise when instruments

are used for data collection, interpretation, and reporting in a manner that is not consistent with their intended use.

Test *reliability* refers to the degree to which a test is consistent and stable in measuring what it is intended to measure. A test is considered reliable if it is consistent within itself and across time. Reliability is usually reported using a coefficient between 0 (no reliability) and 1 (perfect reliability). Generally for the assessments used in a program evaluation .80 and above is considered acceptable. Test validity refers to the degree to which the test actually measures what it claims to measure. Test validity is also the extent to which inferences, conclusions, and decisions made on the basis of test scores are appropriate and meaningful.

Standardized assessments are most commonly used in research designs because they allow for a fair comparison among individual or groups of test takers. They require following a strict protocol so that consistency can be maintained and training the test administrator is a necessary component for this type of assessment. This type of assessment is not usually administered to whole populations, but rather done so on a sampling basis to provide a representative picture of the group. Also, these assessments do not always provide the best information for teachers to use in planning instruction or monitoring individual growth.

Less formal assessments often provide the teacher with more information about the child in a timelier manner to guide instruction; some examples include the Early Learning Scale (Ayers, Stevenson-Boyd, & Frede, 2007) and the Work Sampling System (See Dichtelmiller, Jablon, Dorfman, Marsden, & Meisels, 2000). Progress-monitoring assessments such as observations, checklists, and portfolios provide the teacher with a

systematic, performance-based approach to child assessment that can be used immediately to plan activities and guide instruction. Screening instruments that provide a quick examination of a particular area are useful to alert teachers to an issue in a specific area and diagnostic assessments are usually used individually to identify specific instructional needs once an issue has been identified.

Even when the measure is reliable and valid and being used for its intended purpose there are several criteria that should be considered when assessing young children (Epstein, Schweinhart, DeBruin-Parecki, & Robin, 2004). They are:

1. Assessment should not make children feel anxious or scared;
2. Information should be obtained over time;
3. An attempt should be made to obtain information on the same content area from multiple and diverse sources, especially when repeated instances of data gathering are not feasible;
4. The length of the assessment should be sensitive to young children's interests and attention spans; and
5. Testing for purposes of program accountability should employ appropriate sampling methods whenever feasible.

Research Questions

These questions are addressed in the research designs described in the next section, but this format provides the reader with a quick glance at some of the possible questions to answer through evaluation of the preschool program.

Key effectiveness question:

- What effect does preschool have on children's academic achievement and school skills at kindergarten entry?

Key short-term return on investment question:

- What effect does preschool have on children's grade retention and special education status?

Key longitudinal question:

- How do children who attend preschool compare over time in academic achievement and social skills with children who do not attend the program?

School readiness population questions:

- What is the readiness status of children when they enter kindergarten?
- Is it changing?

Quality questions:

- What is the quality of the preschool programs available to children?
- Is it changing?
- How does the quality of preschool classrooms differ across auspice, provider, or teacher?
- What is the impact of quality of the preschool experience on student outcomes?

Quality improvement questions:

- What is the quality of individual (local agency) preschool programs and how can it be improved?

Workforce questions:

- What are the qualifications of teachers?

- Are they changing?
- What is the impact of teacher degree and teacher pay on quality of preschool?

Dosage questions:

- What is the impact of length of day (or length of year) on the child academic and social outcomes?
- What is the impact of 1 year versus 2 years of preschool experience?

Economic benefits question:

- What is the relationship between the costs of preschool and the (short and/or long term) benefits achieved?

Research Designs

This section provides a description of several options for evaluating a preschool program. The descriptions that follow introduce the design by the research question(s) that can be addressed through the approach, describing the methodology including the sample, data collection and procedures, and analyses, and provide an estimation of cost for each study. A perspective of each research design is offered to provide information regarding the positive and negative aspects of each approach and to highlight how each approach differs.

Five research designs seem mostly likely to be applicable to the evaluation of a preschool program. They may be used separately or combined for greater strength.

1. Nonequivalent groups, post-test using extant data from kindergarten entry and third grade.
2. Nonequivalent groups, post-test only with new data beginning at kindergarten.
3. Nonequivalent groups, pre- and post-test, if there is a waiting list or screen-out.

4. Regression Discontinuity using the birth-date cutoff.
5. Randomized Trial.

1. Utilization of Extant Data: Nonequivalent groups with kindergarten entry and third grade post-test only (Effectiveness, longitudinal, and school readiness questions)

This method relies on a post-hoc examination of state testing data taken at kindergarten entry and/or in third grade, in relation to preschool attendance. This design uses data that is already collected and available.

Groups of students who attended preschool and those who did not attend preschool are identified and compared on these available assessments at these two points in time. This approach provides a relatively quick indication of the impact of preschool by answering the questions, How do children who attend preschool compare at kindergarten entry to children who did not attend the program and how do children who attend preschool compare in third grade? This is a relatively weak design for drawing causal conclusions about the impact of the preschool program. Historically, this design tends to underestimate the effects of preschool programs that target disadvantaged populations.

One method for reducing potential bias is to compare children within districts, perhaps even limiting the participating districts to those that are relatively homogeneous. One could select the districts for participation at random to represent the population of children or districts. Alternatively, one could focus on the districts with the most children served because they reach the largest number of children. Finally, one could select the districts that would best facilitate the collection of data or that have the most complete

and accurate existing data records if there is substantial variation in data quality across districts.

Data would then be compiled on selected students in grade kindergarten and grade 3 at the current time. In addition to their scores at kindergarten entry and in third grade, it is crucial to obtain as accurate information as possible about the student's attendance in the preschool program and length of attendance. Other important information that would be necessary to match the groups for analyses (those students who attended the program and those who did not attend) would include family demographic information such as maternal education level, primary language spoken in the home, and family income level.

Matching groups (or stratifying samples) in this post-hoc design is one approach to control for differences that naturally occur between the groups of students, thereby reducing the potential for selection bias. This means that the two groups would have similar distributions of maternal education, family income level, and look similar in terms of the primary language spoken in the home. There are multiple approaches to matching and matching can be combined with statistical analyses that control for the matching variables. Unfortunately, there is no guarantee that matching will eliminate selection bias and under some conditions (e.g., matching groups that are only partially overlapping) it can increase bias.

One serious issue in this data collection is the accuracy of preschool attendance records. It is necessary to assign students to the groups based on attendance in the preschool program. This is often obtained by parent report either upon entry to kindergarten or in retrospect. Misreporting of preschool attendance is common and can lead one to an underestimate of the impact of preschool education. In addition, accounts

of how long the child attended and particularly which program the student attended may not be available. One cannot be sure that because a child was enrolled in a program that he or she actually received the treatment of preschool without attendance records from the institutions. On the flip side, one cannot be clear about the type of services or experiences the child who was not participating in the preschool program received. It is possible that this child was placed in a high-quality child care environment that provided similar services to the state program. If data can be obtained about the experiences of the comparison group, this situation can be improved. However, this is not often possible. This again can impact the composition of the groups and thus influence the results of the effect.

Another key issue is determining eligibility for children in the comparison group. Ideally, only children who were eligible for preschool, but did not attend preschool should be included in the matched group to the treatment group. However, obtaining eligibility status at the time of preschool would be difficult since these children were in the preschool program four years ago (when looking at current third grade data). An option is to consider children eligible for preschool based on current family characteristics during the third grade year. Inaccurate assignment to the group may occur based on this information because a child that meets the qualification for preschool during the third grade year may or may not have qualified during preschool enrollment and vice-versa. Similar to the issues presented earlier, this inaccurate assignment to a group can cause an underestimation of the effect of preschool.

Yet another consideration is the students who would have been part of the sample during the preschool and kindergarten years because of either eligibility or attendance in

the program that are no longer in the school system. The question is whether this loss creates a type of selection bias in the group. No definitive answer for this question is possible unless the children are located and examined in the study, which increases the cost of the design.

A related issue is that when using third grade data, the study excludes children who were retained in grade or who have been excluded from testing because they are in some group that is not tested (some special education children for example). If the study employs third grade data that are not for the most recent year, then children who were retained can be recovered from the next year's third grade data. However, the question of how to select and match samples becomes complex when it is expected that preschool attendance affects who is retained and, thus, what percentage of preschool and comparison children from a given age cohort should be in each grade level. In addition, the third grade data include students who do not belong in that age cohort because they were previously retained. These children can be excluded by matching on entry year to kindergarten and age. The third grade data also include other children who could not have participated in the preschool program because they moved into the district from elsewhere; these children should be excluded from the sample.

This type of post-hoc analyses using existing data does not provide the evaluator with any pre-treatment data to determine that the two groups were equivalent at the start of the treatment (preschool). Thus, it limits the strength of the research and raises questions of equality in the groups even if they are matched to the best ability based on current data. It is because of this particular issue that this approach is not recommended for high stakes decision making such as whether the program continues or expands. It

does provide the state with an examination of the effects of the preschool program, but underestimation of effects remains a serious concern.

One final caveat with any approach that looks at long term effects is that it provides an examination of the preschool program that was in effect years prior to the study. There is no getting around this problem with either a retrospective or prospective study. However, in this case, it is necessary to consider how interested the state is in the performance of the program as it existed four years earlier, given the level of quality at that point in time and other program characteristics including who was served. So, consideration must be made to the policy implications of a study if changes were subsequently implemented in delivery of services such as eligibility, standards, materials, and teacher qualifications.

The timeframe and cost of this approach to evaluation of a preschool program is dependent on several factors. One factor is the availability of data. The use of existing data within a state system often presents itself as the least costly option for examining the impact of preschool. However, if state data sets are not designed with this intent in mind, the collection and organization of data can quickly raise the cost. If a research team must sift through several data bases for the needed information the time involved and the cost will surely increase. If the research team must retrieve information from students' individual hard copy files the time and cost factors may increase exponentially and this type of study may not be worth the expense that would be incurred.

A second factor to consider is the cooperation of local school districts in providing the required data and offering access to the necessary records. This approach would require strong support from the state officials in order to impress upon school

officials the importance of facilitating the collection of this data in a timely and organized manner. How much local cooperation is required and how much can be done with data collected by the state depends on the type and quality of data available that can be linked to the third grade data, and whether there are unique identifiers that can be used to link other data with test scores.

A final consideration is the number of districts and children to be included in the data set that would be analyzed. The sample is potentially quite large, but if there are issues of incomplete data, and the number of children participating in the preschool program or not participating in the preschool program in a district is quite small, then sample size might become a limitation in some districts.

Once an approach is decided upon based on the available data, selection of the sample and collection of the data (unless there are data sets where the information is already available that can be linked) may take from four to six months dependent upon the difficulty in extracting the necessary data points. The next step would be the entering and cleaning of the data which looks for abnormalities in the grouping, examines for missing data, and creates a data set that is ready for analyses. Finally, statistical analyses would be performed to examine the research question. The entering, cleaning, and analyses of the data could take approximately 2-3 months depending upon the sample size and the condition of the data collected. However, it is not infrequently the case that unanticipated problems are encountered either with the data or the analyses that require more extensive analyses including testing of alternative models and their assumptions.

We estimate the direct cost of this approach to be somewhere between $150,000 and $190,000. For this approach almost all of the cost is in the salaries of the research

team. Thus, the costs could vary greatly if the salaries are markedly different. We based the salaries on reasonable estimations of senior and junior researchers.

Examples of reports using large, general purpose data sets to study preschool, such as the Early Childhood Longitudinal Study (ECLS-K) are presented in Appendix A.

2. Nonequivalent groups, post-test only with new data beginning in kindergarten (Effectiveness, longitudinal, short-term return on investment, and school readiness questions)

This is the most typical approach to the examination of the impact of preschool on academic achievement. This design compares children in the same age cohort who did and did not attend a preschool program (forming a *preschool group* and a *no preschool* group). As with other prospective designs, several key questions can be addressed. First, what effect does preschool have on children's academic achievement and social skills at kindergarten entry? Second, how do children who attend preschool compare over time in academic achievement and social skills with children who do not attend the program? Third, what effect does preschool have on grade retention and special education status?

This approach begins with children in kindergarten. One approach would be to identify a random sample of kindergarten children. This sample would be selected without consideration of preschool participation thus ensuring that a proportionally appropriate number of children would not have attended the preschool program. The children who did not attend the preschool program would form the control group for this study. Statistical controls could be used to adjust for differences between those who attend and do not attend the preschool program. Alternatively, a random sample of those who attended preschool can be selected from lists of children who attended the program

that would then be matched to kindergarten lists or by identifying those children in kindergarten some other way. This preschool sample would then be matched with a comparison group on as many variables as possible (or using some other matching technique such as propensity scores). As discussed earlier, matching can have disadvantages, but it can perform better than statistical adjustments when relationships between the child and family characteristics and outcome measures are nonlinear and not well-understood.

As with any post-test only design, there are several issues that do not arise when there is prospective random assignment. The most significant issue is selection bias that often leads to underestimating the impact of the preschool program. Because we do not have a pre-assessment of the students' academic abilities and social skills before the treatment for the preschool group and at the same time period for the control group we cannot assume the two groups did not differ prior to the treatment.

Beginning to study the students in kindergarten poses the issue of the type of services or experiences the child received during the preschool years for the control group. Again, the comparison children's experiences are unknown unless data are collected from parents about them, and this may not provide accurate information especially about educational quality. These experiences or lack of experiences influence the control group greatly and may influence the results of the impact of preschool. The extent to which this is a problem depends on the degree to which the relevant question is how much the preschool program contributes to child development compared to no program participation or compared to whatever program participation children do or do not receive when the state does not provide a program.

A combination approach using this design with regression-discontinuity design offers some additional protection from the problem of selection bias, as discussed later.

An additional issue with the post-hoc design presented here that is not present in a study that begins at entry to preschool is that the quality of preschool experience that the treatment group received in the preschool program may be difficult to determine. It is possible to do some examination of quality during this year on the preschool programs and match the student to the program. However, issues arise in obtaining accurate reports of which program was attended and also the possibility that the quality in the current year will vary somewhat from the quality in the prior year. This may not be a huge problem, but it is an additional source of error.

Family demographic information should be collected on the selected sample of children including maternal education, family income and primary language spoken in the home (see Appendix B for a description of the content and the procedures). Children's academic achievement is assessed by academic achievement outcomes in language, literacy, and mathematics. Social emotional assessments are completed by the teacher for the students. Children's special education and retention status should be monitored during the study. The assessment instruments change over time because the nature of assessing young children differs from that of examining the academic achievement of older children. Tests are not always best suited for all age groups in the range of early kindergarten through the end of third grade. See Appendix B for a list of recommended assessments.

Child assessments collected at kindergarten entry describes the impact of the preschool program at kindergarten entry. Child assessments at the end of kindergarten, first, second and third grades examine the impact of the preschool program over time.

An issue with this and any other longitudinal approach is that it will take a considerable amount of time to produce estimates of long-term effects. The effects on children at kindergarten entry can be assessed and evaluated within a year of their kindergarten entry. However, estimates of effects at the end of third grade or entry to fourth grade will require another three or four years, at least. Even so, such a study does produce measures of effects on outcomes each year. One must judge whether the program is likely to stay substantially the same over that time, and when the information must be received to be timely.

A final issue is child tracking. It is costly and often difficult to maintain an accurate database of children over a 4 year period. Attrition in the sample occurs because of movement both within the district and outside the district and this often creates a further difference in the preschool versus no preschool groups. However, setting up a clear system at the onset of the study and following the protocol carefully will assist in this necessary aspect of the research.

The estimated costs of this approach vary greatly depending on several factors listed below. The estimated total direct cost, not including indirect costs such as facilities and administration, for year one when the children are in kindergarten is $325,000-$425,000. Costs in subsequent years would be reduced up to $100,000 because children would only be tested once per year.

- We based the salaries on reasonable estimations of senior and junior researchers but these could vary by location. Cost of living variations and geographic spread of the selected sites will affect the expenses as well.
- The sample size of the study will have an impact on the cost.
- There would be an increase cost for a large Spanish speaking population because assessments are then done twice (once in Spanish and once in English).

The direct cost of this approach could be reduced by $100,000 if tests were administered only at the beginning of the kindergarten year and no follow up data were collected. However, this would only answer the question of whether preschool has short term benefits.

3. Pre-and Post-test design with nonequivalent comparison groups (Effectiveness, longitudinal, short-term return on investment, and school readiness questions).

Sometimes it is possible to identify a comparison group at the time of recruitment or enrollment into the preschool program. For example, sometimes programs have waiting lists. The waiting list can be used as a comparison group, matched to a treatment sample, and both groups can be given a pre-test. In other cases, programs may screen applicants for the program to determine eligibility. If any kind of continuous measure (income, a risk index, screening test, or some combination of criteria) is used to determine eligibility, then this information can be used to adjust for differences between the groups in a kind of regression discontinuity analysis (different from that described later). Again, a pre-test can be administered to both groups. Although this design is less often possible than the post-test only design, it offers the great advantage of a potential reduction in selection bias. Thus, it is less likely to underestimate the effects of the

preschool program. The pre-test also substantially increases the statistical power enabling the study to more reliably identify modest program effects. In other respects it shares the strengths and limitations of the longitudinal nonequivalent comparison group design beginning a year later in kindergarten.

The total direct cost of this type of study is very similar to the nonequivalent groups, post-test only, described above, from $325,000 to $425,000 for year one. With one additional year of funding needed to follow children to 3rd grade. All of the same assumptions and issues in estimating these costs outlined above are applicable to this research approach.

Studies using nonequivalent comparison groups, such as the Michigan School Readiness Program, are presented in Appendix A.

4. Regression-Discontinuity Design (RDD) (Effectiveness and school readiness questions)

The regression discontinuity design (RDD) methodology assesses the effects of participation in the preschool initiative on children's skills after one year of the program, typically at entry to the 4 year old preschool year or at kindergarten entry. The RDD approach provides an estimate of short-term outcomes after one year of preschool; it can not be used to estimate long-term effects or to compare the effects of one and two years of program participation, though it can be combined with another longitudinal component that estimates both of these.

The RDD is a statistical approach that addresses the problem of selection bias, which is common to many education studies, and a particularly pernicious problem for preschool programs that are targeted or means-tested. To mitigate selection bias, the

RDD compares two groups of children who enroll voluntarily in the preschool program. Children just entering the preschool program are the control group, since they have voluntarily selected the program (and been selected into it) but have not yet received it. Children who have just finished the preschool program and are currently beginning the 4 year old preschool year or kindergarten are the treatment group, since they also voluntarily selected the program and received the treatment. The RDD methodology utilizes stringent, specified age cut-off for preschool eligibility to define the treatment and control groups among the children in the study. Thus, it is only a possibility where eligibility for the preschool program is subject to a strict birth date cutoff.

One way to think about this design is as essentially randomly assigning children around the birth date cutoff. In the extreme case, this design compares two children who differ only in that one was born the day before the age cutoff (and is currently entering preschool) and the other the day after the age cutoff (and is currently entering the second year of preschool at age 4 or entering kindergarten after completing preschool). Otherwise such children are likely to differ in no systematic ways from each other. When both of these children are tested at the start of the program year, the difference in their scores can provide an unbiased estimate of the effects of the preschool program. The sample size would be very small if only children with these birth dates were included in the study, but this approach is applied to all children in the study by taking into account the proximity of their birth date to the age cutoff. Data are also collected on children's family background as an additional means of ensuring that the treatment and control groups are comparable and increasing statistical power.

Another way of thinking about the RDD approach is that it models selection into the program on a variable that is subject to almost no error (the child's birth date). RDD methodology takes advantage of the state's enrollment policy determined by a child's date of birth by creating two groups of children. One group will have participated in the preschool program at age three or four. The other group will be currently enrolled in and just beginning the preschool program who did not attend the state program at age three. The evaluators can model selection into the program based on the knowledge that there are no other variables likely to affect the child's test scores that vary abruptly at the age cutoff for entry to the preschool program.

The treatment group can be drawn at random from lists of the previous year's preschool participants or from entering kindergarten children as long as their participation can be determined accurately. The sample may be stratified by district, characteristics of the preschool program or auspice. These children will be the *Preschool* group because they have received the "treatment" of preschool the previous year. Three and four-year-old children will be selected who are just enrolling in the same programs. This group of children will be the *No Preschool* group because at the time of assessment, early in the school year, they have not yet received the "treatment" of preschool.

To examine the research question of the effect of preschool on children's academic achievement after one year of preschool children are assessed using academic achievement outcomes in language, literacy, and mathematics. Child assessments must occur during the first few weeks of school for both the kindergarten group and the group just entering preschool. (See Appendix B for a list of recommended assessments.)

Effects on social emotional skills cannot generally be estimated using the RDD approach. This is because the rating scales used to measure social development are generally completed by the teacher and require the teacher to compare the children to typical children this age. In essence, this brings into play another variable that varies precisely with the birthdate cutoff. If a child is one day too young, the teacher compares that child to the typical preschooler. If a child just meets the cutoff, the child's behavior is compared to that of a typical kindergartener.

RDD analysis provides a regression line of the children's predicted test scores by age, measured by the number of days their birth date is from the program enrollment cutoff date. The discontinuity in the line at the cutoff date is the estimated effect of the preschool program. See the diagrams in Appendix C for a pictorial description of these analyses. These analyses will control for student ethnicity, gender, age and school district and take into account the effects of clustering by classroom in the sample.

Identification of children for the study would best be accomplished during the summer months as children enroll in both preschool and kindergarten or in the first week of school, if at all possible. It is best if assessments occur in this first week as well, but it seems reasonable for assessments to occur before the sixth week of school. This is because of the importance of examining children's achievement before the control group receives the treatment of either 3 year old or 4 year old preschool and the treatment group is influenced by the 4 year old or kindergarten year.

The RDD approach requires a substantial amount of methodological sophistication because its success depends on correctly modeling relationships. Thus, a great deal of statistical modeling and testing of assumptions is required. This includes

testing alternative cut points for the discontinuity, testing for effects on theoretically unrelated "outcomes" (where there should be no effect), estimating nonlinear models, conducting nonparametric regressions, and estimating alternative models if some children are misallocated (violate the assignment rule). This requires both considerable expertise and time.

RDD design is an excellent option when it is not possible to employ the randomized trial approach to research the effects of the preschool program on children's academic outcomes. This approach also eliminates the need to find a control group of children that did not receive the state's preschool program, which can be costly and time consuming. RDD provides a clear manner in examining the effects of the program at kindergarten entry without the complications of random assignment and without the issue of selection bias. However, the RDD approach cannot provide an estimate of effects beyond kindergarten entry because if it was employed a year later, it would provide an estimate of the added effects of kindergarten. Therefore, this design is best when coupled with one of the longitudinal designs described earlier.

RDD cannot be used to estimate the effects of the preschool program beyond one year. For example, if applied in first grade, it would estimate the effects of kindergarten. However, if a longitudinal nonequivalent comparison group design as discussed above is also employed, beginning in preschool, the RDD design can be used to estimate the amount and direction of selection bias. If the RDD estimates and nonequivalent comparison group design estimates for effects at entry to the four year old program or kindergarten are highly similar then one can conclude that selection bias is not a serious problem in the latter approach. If they differ substantially, then selection bias must be

considered a serious problem. It may be possible to test alternative models with the latter data to find one that closely approximates the results of the RDD approach. This model could then be used going forward in the longitudinal study. However, such success likely depends on having a relatively rich set of measures of the children and their families, and is not guaranteed. If a better model cannot be produced, at least the magnitude and direction of selection bias can be identified. This may or may not be satisfactory.

The RDD approach employing the birth date cutoff has been used by Dr. William Gormley and colleagues in studies estimating the effects of Oklahoma's universal preschool education program in the Tulsa school district (Gormley, Gayer, Phillips, & Dawson, 2005) and by NIEER to estimate the effects of programs in Oklahoma and other states (Barnett, Howes, & Jung, 2008; Barnett & Massey, 2007; Wong, Cook, Barnett, & Jung, 2008). (See Appendix A for more information about these studies and others using RDD.)

One of NIEER's studies of New Jersey's Abbott preschool program estimated effects of the 3 year old and the 4 year old programs. From these studies we have learned a number of valuable practical lessons about what appears to strengthen the results of an RDD study. First, larger sample sizes produce more stable estimates across various functional forms (e.g., linear, quadratic, cubic) and other tests of the model making it easier to identify the most appropriate model for the analysis, a crucial issue for correctly estimating program effects. A sample size of 3000 (1500 in each group) is not overkill. Second, it is important that both preschool and kindergarten samples be representative and that there has not been any significant change in the population served between the two years. Third, children should be assessed as early in the school year as possible. In

Tulsa, testing has been done successfully during the first week of school. Fourth, data on the quality of the preschool programs children attended can help make a more persuasive case that the programs produced the estimated effects. Fifth, the more data collected on child and family characteristics, the better. At a minimum it is desirable to have information on each child's ethnicity, home language, income, and location (which preschool and school attended). In addition, it would be useful to have information on maternal education level and even on attitudes and parenting practices.

The estimated direct cost for a RDD study ranges from $260,000 to $350,000. This range again reflects different assumptions about sample size, instrumentation, and cost of living. Adding the longitudinal nonequivalent comparison group design would increase the cost in year one by approximately $50,000-$150,000. Costs of subsequent years would remain mainly the same.

5. Randomized Trial (Effectiveness, longitudinal, short-term return on investment, and school readiness questions).

A randomized trial is the best method for answering well-defined questions about "what works" (Feuer, Towne, & Shavelson, 2002). It calls for random assignment to the preschool program and creates a clean, unbiased sample of children that attend preschool and children who do not attend preschool. This approach works best when acceptance to the program is done by lottery system. In some circumstances, all possible children eligible and interested in attending preschool are grouped and then a sample is selected randomly to attend the preschool program. In others, randomization is applied only to some group at the margin (for example, the last 20 applicants at each location, children

who are above 150% of the poverty line, or those on a waiting list). A longitudinal approach is then implemented by following the two groups over time.

A randomized trial approach provides the greatest confidence that the treatment and comparison groups do not differ on either measured or unmeasured characteristics. It also provides the greatest statistical power for any given sample size. Like the other models, it is strengthened by a pre-test, but a pre-test is not necessary. It would offer the most reliable answers to several basic questions: (1) What effect does the preschool program have on children's academic achievement and social skills at kindergarten entry? (2) How do children who attend the preschool program compare over time in academic achievement and social skills with children who do not attend the program? (3) What effect does the preschool program have on grade retention and special education status?

The randomized trial research design requires that all children at age three that are eligible and interested in attending preschool at age four are grouped together. At this time, family demographics are collected such as maternal education, family income, and primary language spoken in the home (see Appendix B for further data points that can be collected). Then, a sample, stratified on family characteristics, is selected randomly from this universe to attend the preschool program the following year when the children are four-years-old.

Those in the selected group are the treatment group and the remaining children on the list form the control group or *no preschool* group. Both groups are followed over time beginning immediately. Assignment to attend preschool for only one year at age four or for two years beginning at age three can occur at this time if there is interest in the

supplementary research question, what is the impact of two years of preschool versus one year of preschool? Regardless of this decision, it is necessary to track the children during these two years to consider what type, if any, preschool experiences they are participating in and to collect baseline data.

This information about the child's experiences at age three for the treatment group (age four treatment would be the preschool program) and age three and four for the control group should be considered in the analyses and reporting of the results. Formal assessments of both groups of children should be conducted fall and spring of the four-year-old preschool year, fall and spring of the kindergarten year, and then spring of first grade, second grade, and third grade. Additionally, a third grade standardized achievement test can be included in the analyses.

Children's academic achievement is assessed by academic achievement outcomes in language, literacy, and mathematics. Social emotional assessments are completed by the teacher for the students. The assessment instruments will change over time because the nature of assessing young children differs from that of examining the academic achievement of older children. Tests are not always best suited for all age groups in the range of early kindergarten through the end of third grade. (See Appendix B for a list of recommended assessments.) Children's special education and retention status would be monitored during the years of the study. Studies using randomized trial, such as the Head Start Research, are presented in Appendix A.

The direct cost of the randomized trial ranges from $350,000 to $700,000 for year one, with four subsequent years of funding at similar levels. As above, there will be an increased cost for a large Spanish speaking population. Additionally, sample size, cost of

living and spread of sample sites will affect cost. The estimated costs are direct costs only and do not include indirect costs such as facilities and administration.

Examination of Quality (Quality and quality improvement questions).

The questions considering quality may be added to any of the above research studies that can accurately link program data to specific children, and is best addressed by studies with pre-test data. If the there is interest in summarizing the quality of the program as well as examining the impact that quality has on the effect of preschool then examination of the quality of a sample of the preschool classrooms is necessary. This classroom observation piece can add a considerable cost to the research study, but we feel that it is a necessary component to truly evaluate the preschool program.

The quality of the program is generally determined by observation by outside observers trained to reliability on a battery of classroom observation instruments that examine several aspects of the classroom. Some characteristics that are evaluated include general space and furnishings, personal care routines, various activities in several areas (e.g. fine motor, art, dramatic play), interaction between the children and the teachers, language and reasoning materials and concepts, mathematical materials and concepts, program structure, and provisions for parents and staff. (See Appendix B for a description of recommended observation tools.)

By including this classroom-level aspect of an evaluation information is summarized to examine such questions as: How does quality of preschool differ across auspice, provider or teacher? and What is the impact of quality of the preschool experience on student outcomes?

The randomized trial approach can offer the most reliable data to answer the second question regarding the impact of quality on student outcomes. However, it provides a better answer to this question only if children within the treatment group are randomly assigned to programs of different quality. This is not often the case. If random assignment within the treatment group does not occur, it is unclear how much better the answer to the quality question is from what is provided by other approaches. Again, the randomized trial only has a clear advantage for these other questions if children are randomly assigned within the program.

Additional Questions of Interest (Workforce and dosage questions)

There may also be interest in the following additional research questions:

What is the impact of teacher degree and teacher pay on the quality of preschool? This data can be obtained through a workforce registry, but would require the connection between student and teacher as an additional data point. This also requires matching students in the groups or randomly assigning students within the program using either the pre-and post-test design with nonequivalent comparison groups, the RDD, or the randomized trial.

Consideration of the impact of the length of day on the child academic and social emotional outcomes can be studied at a given time such as kindergarten entry. This also requires matching students in the groups or randomly assigning students within the program using the pre-and post-test design with nonequivalent comparison groups, the RDD, or the randomized trial.

Examination of the question of 1 year of preschool versus 2 years can be done but requires good data on children's attendance at age 3 and 4 and complete data on child and

family characteristics so that these factors can be controlled for in the analyses. This comparison can be added to any of the longitudinal designs and as a non-equivalent comparison in the RDD if the information is available.

The quality of kindergarten through grade three classrooms influence on the effect of preschool may be of interest. This requires classroom quality examinations at these grade levels, random assignment to classrooms or matching procedures, and a system to connect the child to the kindergarten through third grade teachers.

Lastly, the preschool teacher's proficiency in language (either Spanish or English) may influence the effect of preschool on child outcomes. There may be interest in examining the teacher's vocabularies using the PPVT or TVIP (See Appendix B for a description of these instruments.).

Economic Analysis

Cost Analysis

Any economic analysis has to begin by specifying the cost of that investment. This typically requires some kind of cost analysis rather than simply taking some figure from a preschool program budget (Barnett, Frede, Cox, & Black, 1994). The reason for this is that no single budget has all of the information about a program's cost. State administrative costs are usually not in the same budget as state funding for direct services. Infrastructure costs often are in separate budgets, as are capital costs. Most preschool programs are partnerships with local agencies or private providers and even if there is no explicit requirement for cost sharing, local public school and other local government resources and private resources (donated space, United Way funds) including parent fees may contribute to covering program costs. Costs may not appear in any

current budget because they are in the past (facilities that were built long ago) or the future (retirement pay and health care for current employees).

Two approaches can be taken to cost estimation. One is to try to identify all of the sources of revenue and in-kind resources and identify their value. The other is to identify all of ingredients and then cost them out. How much effort this requires depends on whether all of the data are collected from a sample of programs or whether simple assumptions are used to ballpark various aspects of cost. It is important to keep in mind that most of the cost is accounted for by direct service staff.

Cost-benefit Analysis

An economic analysis can be conducted through a cost-benefit analysis (CBA). The goal of a CBA is to translate all costs and as many of the effects as possible into a common measure—money. A basic guide is provided by Levin and McEwan (2001). The greatest problem for cost-benefit analysis is that it can be very difficult, if not impossible, to produce credible measures of the monetary value of some outcomes. The field has pushed this quite far in CBA of early childhood education programs, however. Also, even a partial monetization of the benefits can be highly illuminating for preschool programs. A more serious limitation for most programs is that most of benefits of preschool education for which a dollar value can be estimated occur far into the future. The best way to estimate these is a prospective longitudinal study. However, it takes seven years for children exiting a preschool program to make it through elementary school and 13 for them to finish high school. That means waiting a long time for study results. An alternative is to try to work backwards from children who have just completed high school, though this means that the program being evaluated is over a

decade old. An intermediate strategy is to estimate results for a synthetic cohort that will reflect the results of current and past programs, though we do not know of anyone who has taken this approach. Finally, it is possible to estimate benefits from other studies that have already been completed by making assumptions that link the results of these studies to a specific state's preschool education programs and the population it serves.

The approach most frequently adopted in studies of state programs has been to estimate benefits by relying on existing longitudinal research—the Perry Preschool, Abecedarian, and Chicago Child-Parent Center (CPC) studies (Barnett, 1993; Barnett & Masse, 2007; Belfield, Nores, Barnett, & Schweinhart, 2006; Reynolds, Temple, Robertson, & Mann, 2002; Temple & Reynolds, 2007). The Perry and CPC studies are most often relied upon because their programs are most similar to typical preschool programs, for example, both are half-day programs. However, few state programs are as intensive as these programs, with the CPC being closest to typical practice. In states where some programs are full-day year-round, it may be necessary to triangulate across all three studies to come up with useful estimates for benefits. Examples of various approaches are available as a guide (Belfield, 2005, 2006a, 2006b, 2006c; Karoly, Kilburn, & Cannon, 2005; Lynch, 2004).

The process of estimating benefits can range from quite simple to highly complex. For example, one could simply extrapolate from one of the studies by assuming that benefits in a state were proportional to relative program effectiveness. Thus, if the state program were found to have initial effects or effects into the early grades on achievement that were about ¾ of the size of those for the CPC program, then it might be assumed that ¾ of all of the CPC benefits would obtain for that state. A more complex approach

would estimate the value of effects 3/4 the size of those in CPC on each individual benefit. For example, one could estimate the effects of a 5 percentage point reduction in special education on a state's special education costs, a 10 percentage point reduction in high school dropout on earnings and tax revenue, and a 7 percentage point reduction in arrest rates on the costs of crime and the criminal justice system. (Note these percentages were selected at random for illustrative purposes.)

The process of estimating benefits can be even more involved than indicated above. Belfield (2006d) has identified additional potential benefits beyond those estimated in the three CBA studies cited earlier, and these can be estimated from estimates of program effects. These and other benefit estimates will be more credible, the more actual data on program effects underlie the program benefit estimates. Thus, the least confidence can be placed in benefit estimates that are based only on estimates of the program's cognitive benefits. Greater confidence is gained if program effect estimates include effects on social and emotional developments well. Confidence also increases if there are estimates of later impacts on achievement and behavior in the early grades, grade repetition, and special education placements. Nevertheless, even rough calculations of benefits can be useful. In every analysis, it is appropriate to perform sensitivity analyses to demonstrate how much benefit estimates vary with changes in assumptions, some of which may be highly uncertain. Finally, the rate of return on investment is calculated by estimating the interest rate that would generate an equivalent stream of benefits over time. This makes sense only if there were a fairly complete estimate of the benefit stream over many years. Otherwise the economic analysis will

seriously underestimate the return, and unlike a partial benefit, it cannot simply be subtracted from cost as a way to judge value of the partial benefits.

Cost and Economic Analysis for Decision Making

Cost and benefit information can contribute to a wide array of internal and external decisions. Examples are provided below of the types of decisions for which this information can be important.

1. Planning budget allocations and projecting resource needs. Average and marginal cost information are required in order to know how changes in the number and types of children and families to be served will affect the needs of various programs for resources. In planning a program expansion, how much money will be needed for new facilities? How much for personnel preparation? How much for annual operating costs?

2. Improving the efficiency of program operation. Private sector providers who depend on fees have more information than public sector providers about the efficiency of program operation. Private providers must satisfy their customers with quality and price or they lose clients. Because they have less information, public providers may find cost analysis even more useful than private providers. Both types of providers can use cost analysis to try to identify modifications of program ingredients or activities that would save money without reducing benefits or that would generate more or better services without increasing cost.

3. Setting fair and adequate fee or payment schedules. Many programs charge fees for their services. A program or agency may be interested in this because it charges for the services or because it reimburses other agencies for services they provide

for its clients. They may need to know how costs vary with the population served or service area, as well.

4. Identifying the impacts of and finding the best ways to meet regulatory and licensing standards. Agencies that regulate and license education programs need to be sensitive to the effects of their regulations and licensing requirements on costs. Some regulations and standards may have little or no impact on costs. Others may have quite large impacts.

5. Influencing decisions and judgments made by people external to the program. Cost analysis provides a basis for justifying requests for larger budgets to maintain or improve an existing program or to expand by providing a new program. ROI analysis provides a basis for arguments about the value of the program relative to cost and other public expenditures.

Summary

The designs outlined in this paper provide guidance for preschool evaluation and offer various approaches based on the intention or research questions, data availability, and budget. Although these research designs are separated into distinct categories, it is often the case that studies do not fit neatly into a particular design but rather are a combination of designs. Here we will discuss several evaluations of preschool that utilized the designs presented in this paper. We will describe what the study examined, the design or designs the study utilized, and the strengths and weaknesses of the approach. Gilliam and Zigler (2000; 2004) have reviewed and reported on state preschool evaluation studies in more depth and the reader is directed to these papers for a more comprehensive review of these types of evaluations.

Utilizing extant data, the state of Tennessee commissioned an outside evaluation of its prekindergarten program (Strategic Research Group, 2008) comparing students who attended the program to those who did not on various student outcome measures already in place. This secondary data analysis used post-hoc non-equivalent comparison group design using a post-test only (NGPO). The research group individually matched students from the program group to the non-prekindergarten group by race, gender, free and reduced lunch status, and school level then the district level when a school level match was not possible.

Inherent in this design is bias where, although the researchers attempted to provide a matched group, the groups may be different at the start of the project. There is often a difference in families that choose to enroll their child in preschool from those that choose not to and, additionally, we cannot account for the experiences that the comparison group had as preschoolers. Thus, especially without pre-test data, we cannot consider the groups equivalent. Often this bias tends to underestimate the effect of preschool and should be considered in interpreting the results of this study.

Due to the late identification of some groups of prekindergarten attendees, at times as late as 5th grade, we cannot be sure that this sample is representative of the population that did attend the program. There may be patterns of testing and data loss that contributes to this. The matched group is also identified within grade level and this often leads to issues of children who have been retained affecting the results. Specifically, within grade level analysis would exclude children who were retained that belong in the age cohort and include those who were retained and do not belong in the

age cohort. This problem becomes worse with each grade higher because grade repetition is cumulative.

Suggestions for improvement in this study include redrawing the comparison samples using a prospective approach or, at the very least, matching groups on age cohort rather than within grade level. Another option is to include another approach such as the Regression Discontinuity Design (RDD) which would allow a check on the extent of bias in the estimated effects at kindergarten. Although the RDD approach cannot be used to assess effects beyond kindergarten, it can tell you how biased your longitudinal study is likely to be.

The New Jersey Abbott Preschool Program Longitudinal Effects Study (APPLES; Frede, Barnett, Jung, Lamy, Figueras, 2007), a study that investigates the educational effects of state preschool as a result of the New Jersey Supreme Court school-funding case *Abbott v. Burke,* also utilizes a longitudinal non-equivalent post-test only (NGPO) approach. However, there are several differences between this research and that conducted for the state of Tennessee. First, the researchers identify the comparison group prospectively at the start of the kindergarten year. This allows for accurate tracking of the students in this age cohort that are retained, skip a grade or move thus keeping the comparison and treatment groups comparable. Second, this research includes RDD in addition to the NGPO. This combination provides a check of the bias in estimated effects at kindergarten in the NGPO approach which allows more confidence in the accuracy of the estimate of effects of preschool. In this particular study it was found that the estimated effects in the longitudinal study were underestimated. So, as the longitudinal study continues one must interpret the results with this understanding. Lastly, the

APPLES study included an examination of classroom quality which was not included in the Tennessee evaluation. The inclusion of this data allows the state to consider the role of quality in interpreting the effect of preschool on child outcomes, if funding and resources allows, and provides a look at the change of quality over time.

While the APPLES study utilized RDD to strengthen the NGPO approach, a study evaluating South Carolina's public preschool program (Frede & Barnett, 1992) utilized a pre- and post-test to enhance the NGPO approach. This study was one component of a larger evaluation of this statewide program for at risk four-year-old children (Barnett, Frede, Mobasher, & Mohr, 1987). Classrooms for participation were selected to provide geographic diversity and waiting lists to be used to form the comparison group. Although a practical approach, this may not have provided an adequate representation of the state's programs. Children were pre-tested upon request for enrollment and placed either in the program or on a waiting list. This formed the two comparison groups prospectively and provided pre-test information. In addition, data were collected on the comparison group's preschool experiences to help explain the counterfactual. Post-tests were conducted at kindergarten and first grade entry. These assessments were conducted by the teacher rather than an outside assessor which can be considered an issue. To provide strength to the study design a measure of program implementation of the High/Scope curriculum was conducted to assure that inadequate implementation was not diluting the treatment.

Similar concerns about generalizability with the South Carolina research described above come to light with the RDD study of Oklahoma's universal prekindergarten program (Gormley et al., 2005). Only one site was utilized and Tulsa

was chosen because it was the largest district in the state, was racially and ethnically diverse, administered tests at points of interest, and granted permission to add additional assessments. Although the authors cannot present these findings as applicable to the entire state, they still produce strong and meaningful results for this district because of the strong design approach. Children from this district were assessed during the first week of school by trained teachers to provide the sample for the RDD. A strict age cut-off date provides a treatment and control group without the influences of selection bias and the early testing offers a look at child outcomes without the confounding influence of the current school year. The study reports results of prekindergarten on short-term cognitive development for children. As highlighted above, using the RDD approach does not allow for longitudinal reporting of results without being coupled with another design.

The gold standard of designs that provides good control of selection bias and provides a longitudinal look at effects is the randomized trial as used by the landmark study, High/Scope's Perry Preschool Study (Schweinhart, Montie, Xiang, Barnett, Belfield, & Nores, 2005). This study began in the 1960's when a sample of low-income African-American children at risk of school failure were randomly assigned to receive high-quality preschool or to a no preschool program group. This research utilized the strongest research design, followed the sample and reported results through age 40, and examined a wide range of domains including education, economic performance, crime, family relationships, and health. In addition, the study reports minimal attrition. With these characteristics this study provides a strong confidence in the results that can be attributed to the effects of preschool.

Each study outlined above has strengths and weaknesses that are inherent in the design. However, regardless of the design used to evaluate preschool, safeguards must be put into place to protect the strength of the evaluation. First, researchers must carefully consider the quality and appropriateness of assessments used in the design. Second, treatment fidelity is a consideration because if the quality of the preschool is substandard or if curriculum is implemented poorly this will impact the results. It is recommended that evaluation of a program or curriculum is implemented only once an acceptable level of quality or implementation is attained. Lastly, poorly thought out or improperly implemented methods or analyses can discount results. Thus, accurate data collection and appropriate and careful analyses of data must be conducted to assure confidence in the results regardless of the quality of the design.

Appendix A

Research Study Examples

1. True experiments with model or small-scale public programs and long-term follow-up

Study Name and References	Pop Served*	Age
Abecedarian Program- 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; Barnett, W. S., & Masse, L. N. (2007). Early childhood program design and economic returns: Comparative benefit-cost analysis of the Abecedarian program and policy implications, *Economics of Education Review*, *26*, 113-125.. For more information see http://www.fpg.unc.edu/~abc/	Disadv	<1 thru 4
Brigham Young University-Larsen, J. & Robinson, C. C. (1989). Later effects of preschool on low risk children. *Early Childhood Research Quarterly*, 4, 133-144.	Adv	3 & 4
CARE-Wasik, B. H., Ramey, C. T., Bryant, D. M., & Sparling, J. J. (1990). A longitudinal study of two early intervention strategies: Project CARE. *Child Development, 61*(6), 1682-1696. EJ 426 160. For more information see http://www.fpg.unc.edu	Disadv	<1 thru 4
Consortium for Longitudinal Studies-Consortium for Longitudinal Studies (1983). *As the twig is bent… Lasting effects of preschool programs*. Hillsdale, NJ: Lawrence Erlbaum.	Disadv	3 & 4
High Scope Curriculum-Schweinhart, L.J. & Weikart, D.P. (1997). The High/Scope preschool curriculum comparison study through age 23. *Early Childhood Research Quarterly, 12*(2), 117-143. Available at http://www.highscope.org/Research/homepage.htm	Disadv	3 & 4
Houston Parent Child Development Center-Johnson, D. & Walker, T. (1991). A follow-up evaluation of the Houston Parent Child Development Center: School performance. *Journal of Early Intervention*, 15(3), 226-236.	Disadv	1-3
Infant Health and Development Program (IHDP) McCormick, M.C., Brooks-Gunn, J., Buka, S.L., Goldman, J., Yu, J., Salganik, M., Scott, D.T., Bennett, F.C., Kay, L.L., Bernbaum, C., Bauer, C.R., Martin, C., Woods, E.R., Martin, A., & Casey, P.H. (2006). Early intervention in low birth weight premature infants: Results at 18 years of age for the Infant Health and Development Program. *Pediatrics, 117*, 771-780. Available at http://www.pediatrics.org/cgi/content/full/117/3/771.	LBW	<1 thru 2
Mauritius Study-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.	Disadv	3 & 4
Milwaukee Project-Ga rber, H.L. (1988). *The Milwaukee Project: Prevention of mental retardation in children at risk.* Washington, DC: American Association on Mental Retardation.	Disadv	<1 thru 4
Perry Preschool Program-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. Available at http://www.highscope.org/Research/PerryProject/perrymain.htm	Disadv	3 & 4

2. True experiments with large-scale public programs

Study Name and References	Pop Served*	Age
Early Head Start-Love, J. M., Kisker, E.E., Ross, C. M., Schochet, P.Z., Brooks-Gunn, J., Paulsell, D., Boller, K., Constantine, J., Vogel, C., Fuligni, A. S., & Brady-Smith, C. (2002/2004). *Making a difference in the lives of infants and toddlers and their families: The impacts of Early Head Start. Volume I: Final technical report.* Princeton, NJ: Mathematica Policy Research Inc. Available at http://www.mathematica-mpr.com/publications/pdfs/ehsfinalvol1.pdf	Disadv	0-3
Even Start (Family Literacy)-St. Pierre, R.G., Layzer, J.I. & Barnes, H.V. (1998). Regenerating two-generation programs. In W.S. Barnett & S.S. Boocock (Eds.) *Early care and education for children in poverty: Promises, programs, and long-term results,* (pp.99-121), Albany, NY: SUNY Press.	Disadv	Wide range
Head Start-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.	Disadv	3-4
Head Start National Impact Study-Puma, M., Bell, S., Cook, R., Heid, C., Lopez, M., Zill, N., Shapiro, G., Broene, P., Mekos, D., Rohacek, M., Quinn, L., Adams, G., Freidman, J. & Bernstein, H. (2005). *Head Start impact study: First year findings.* Washington, DC: US Dept. of Health and Human Services, Administration for Children and Families. Available at http://www.acf.hhs.gov/programs/opre/hs/impact_study/reports/first_yr_finds/first_yr_finds.pdf	Disadv	3-4

3. Quasi-experiments

Study Name and References	Pop Served*	Age
Chicago Child Parent Centers -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.* (Discussion Paper no. 1245-02). Madison, WI: Institute for Research on Poverty.	Disadv	3 & 4
Michigan School Readiness Program-Xiang, Z. & Schweinhart, L. (2002). Effects five years later: The Michigan School Readiness Program evaluation through age 10. High/Scope Educational Research Foundation. Available at http://www.highscope.org/Research/MsrpEvaluation/msrp-Age10-2.pdf	Disadv	3 & 4
NIEER 5 State Study-Barnett, W.S., Lamy, C., & Jung, K. (2005). The effects of state prekindergarten programs on young children's school readiness in five states. Retrieved February 15, 2006 from http://nieer.org/docs/index.php?DocID=129.	All/ Disadv	4
NY Experimental Pre-K-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.	Disadv	4
South Carolina-Barnett, W.S., Frede, E.C., Mobasher, H., & Mohr, P. (1987). The efficacy of public preschool programs and the relationship of program quality to efficacy. *Educational Evaluation and Policy Analysis, 10*(1), 37-49; 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).*	Disadv	4
Tulsa Oklahoma Study-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. Available at http://content.apa.org/journals/dev/41/6/872	All	4

4. Large statistical studies designed to study pre -k where program is observed

Study Name and References	Pop*	Age
Cost, Quality, and Outcomes Study-Peisner-Feinberg, E., Burchinal, M., Clifford, R., Yazejian, N., Culkin, M., Zelazo, J., Howes, C., Byler, P., Kagan, S., & Rustici, J. (1999). *The children of the Cost, Quality, and Outcomes Study go to school.* Chapel Hill, NC: University of North Carolina, Frank Porter Graham Child Development Center.	All	4
Effective Provision of Pre -School Education (EPPE) Project-Sylva, K., Melhuish, E., Sammons, P., Siraj-Blatchford, I., Taggart, B. (2004). *The final report: Effective pre-school education. Technical paper 12*. London: Institute of Education, University of London.	All	3 & 4
NICHD-Vandell, D. L. (2004). Early child care: The known and the unknown. *Merrill-Palmer Quarterly, 50*(3), 387-414.	All	1 thru 4
Northern Ireland-Melhuish, E., Quinn, L., Hanna, K., Sylva, K., Sammons, P., Siraj-Blatchford, I., & Taggart, B. (2006). *Effective pre-school provision in Northern Ireland (EPPNI) Summary report. No. 41*. Department of Education: Northern Ireland Statistics & Research Agency.	All	3 & 4

5. Large, general purpose data sets used to study preschool. Articles listed are examples.

Study Name and References	Pop*	Age
ECLS-K-Denton, K.L., West.J, & Reaney, L.M.(2001). *The kindergarten year: Findings from the Early Childhood Longitudinal Study, Kindergarten class of 1998-99.* NCES 2001-023. Washington, DC: National Center for Educational Statistics; Magnuson, K., Meyers, M., Ruhm, C., & Waldfogel, J. (2004). *Inequality in preschool education and school readiness.* American Education Research Journal, 41, 115-157.	All	3 & 4
National Education Longitudinal Study-Ludwig, J. & Miller, D.L. (2005). *Does Head Start improve children's life chances? Evidence from a regression discontinuity design.* University of California -Davis. Available at http://www.econ.ucdavis.edu/working_papers/05-34.pdf; Fukahori, S. (2000). *The long term effects of Project Head Start: A national-scale longitudinal study*. Available at http://digitalcommons.libraries.columbia.edu/dissertations/AAI9970194/ .	Disadv	3 & 4
National Longitudinal Study-Currie, J. & Neidell, M. (2007). Getting inside the "black box" of Head Start quality: What matters and what doesn't. *Economics of Education Review, 26,* 83-99; Aughinbaugh, A. (2001). Does Head Start yield long-term benefits? *Journal of Human Resources, 36*(4), 641-665.	Disadv	3 & 4
Panel Study on Income Dynamics-Garces, E., Thomas, D., & Currie, J. (2002). Longer-term effects of Head Start. *American Economic Review, 92*(4), 999-1012.	Disadv	3 & 4

**Population abbreviations-Disadv-Disadvantaged, Adv-Advantaged, LBW- Low birth weight*

Appendix B

Instrumentation

Child and Family Characteristics

Data both on child and family characteristics of the sample children are necessary for matching samples. This information is generally collected through short family interviews done at the time of enrollment or at another time by phone. The interview is conducted in the family's home language. The data that are essential include the following:

- Maternal education level;
- Primary language spoken in the home;
- Family income level; and
- Confirmation of preschool attendance status already collected from school records (if applicable).

Additional data that can be collected, but often have limited utility in analyses include information about:

- The child's health and hospital stays;
- The child's dental care;
- Number of siblings;
- Extended family members such as step members;
- Parent involvement in school such as attending parent-teacher conferences; and
- How many times the family has moved.

Classroom Observation Instrumentation

Early Childhood Environment Rating Scale - Revised (ECERS-R; Harms, Clifford, & Cryer, 2005)

Overall program quality is assessed by trained observers using this standardized measure of preschool classroom structure and process. This measure has been used extensively in the field and has well-established validity and reliability. The validity of the measure is supported by high correlations between both the scale items and ratings of items as highly important by a panel of nationally recognized experts, and between scale scores and ratings of classroom quality by experts. Internal consistency as measured by Cronbach's alpha is reported by the authors to be adequate, ranging from .81 to .91. Classroom quality is rated on a 7-point Likert scale, indicating a range of quality from inadequate (1) to excellent (7). The seven ECERS-R subscales are as follows: 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. This instrument provides an excellent look at what the quality of classrooms is in a program.

The Supports for Early Literacy Assessment (SELA; Smith, Davidson & Weisenfeld, 2001)

The extent to which the classroom environment is supportive of children's literacy development is measured with the SELA. This measure is revised with the deletion of 4 items that overlap with the ECERS-R. The revised measure includes 16 items on a scale from 1 to 5, low quality (1) to high quality (5) for the support of early literacy development. Six subscales are: The Literate Environment, Language

Development, Knowledge of Print/Book Concepts, Phonological Awareness, Letters and Words, and Parent Involvement. This instrument is a good indicator of the quality of literacy in the classroom and can be conducted at the same time as the ECERS-R to provide a more complete picture of the classroom.

The Preschool Classroom Mathematics Inventory (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 low quality (1) to high quality (5), and has two subscales, Materials and Numeracy and Other Mathematical Concepts. Internal consistency among the test items as measured by Cronbach's alpha is good at .86. The PCMI has been found to predict child progress on a standardized math assessment (Frede, Lamy, & Boyd, 2005). This instrument adds to the value of the ECERS-R and SELA by examining mathematics, an area often neglected by preschool teachers, and data for this instrument can also be collected during the same observation time with the two previous instruments.

Snapshot (Ritchie, Howes, Kraft-Sayre, & Weiser, 2002)

This observation tool measures how children and teachers spend their time in the classroom. Used in conjunction with global measures of classroom quality in national studies, the Snapshot has been shown to predict child progress. The Snapshot has good

inter-observer reliability, with a kappa value of .95. (Pianta, Howes, Burchinal, Bryant, Clifford, Early, and Barbarin, 2005)

Classroom Assessment Scoring System (CLASS; Pianta, LaParo, & Hamre, 2006)

This is an observational system that assesses classroom practices in preschool through third grade by measuring the interactions between students and adults. These practices are broadly grouped across three domains of quality of instruction, social/emotional climate and classroom management. The instrument reports convergent validity demonstrated by a relationship between the CLASS and the ECERS and sufficient reliability was reported by internal consistency of the scales that make up two factors in the CLASS with alphas of .85 and .88 (LaParo, Pianta, & Stuhlman, 2004).

Classroom Observation Training

Trained and reliable observers are necessary for the classroom observation of quality. Initial training in administering the observation protocol that includes the *ECERS-R, SELA, and PCMI* takes place in two full day workshops. Trainees then observe in classrooms alongside a trained observer to establish reliability on each observation instrument. The scores of the trainee and the reliable observer are then compared, item by item. The true score for each item is determined through discussion but is generally that of the trained observer. A reliability score for the trainee is computed by determining how many exact matches by item she/he has with the true score and how many are only one point above or below the true score. For the ECERS-R, the trainee must complete three observations with 80% or above exact matches or one-away from the true score and no less than 65% exact agreement. The trainee must achieve 70% exact agreement for the PCMI and SELA for all three sessions. After five sessions, if the

observer is not reliable, he or she is not included in data collection. Shadow scoring is repeated every six weeks.

Initial training in administering the CLASS takes place in two full-day workshops. Trainees must establish reliability by viewing a video on day three. The codes of the trainee are then compared with the master codes, item by item. Master codes have been previously established by measure authors for each video segment. A reliability percentage for the trainee is computed by determining how many codes are within 1 of each master code. The trainee must be within one on 80% of the master codes across five 20-minute video segments.

Training in administering the Snapshot also begins with two full-day workshops, followed by a practice day in preschool classrooms alongside a trained observer. Reliability is established on video on day four. Master video codes have been predetermined by measure authors. Trainees must achieve 75% exact agreement with master codes across four 20-minute video segments. Percentages of accuracy are calculated for each code and anecdotal feedback is provided upon completion of reliability.

In the case that an observer does not achieve reliability criteria during the first session, a second reliability session is available for both CLASS and Snapshot. This typically takes place about a week later, after further preparation and consultation on the measure.

Child Outcome Assessments

After reviewing methods and instruments used in state and local school readiness evaluations, Brown, Scott-Little, Amwake, and Wynn (2007) report several

recommendations to guide the selection and implementation of child assessments. They are: (1) select outcomes for assessment that match the goals of the program and address the components of children's learning and development that are linked with later school success, (2) define the purpose and select the instruments based on that purpose, (3) select instruments that have been successful with similar children to your sample, (4) select culturally and linguistically appropriate instruments for the children who will be assessed, and (5) determine if an outside assessor or someone who works directly with the child would be the best collector of the data.

It is also necessary to utilize instruments that demonstrate reliability and validity, as described earlier in this document, and are developmentally appropriate for the age-group for which it is administered. Children's receptive vocabulary, emergent literacy, early math skills and social skills should be assessed with a battery of instruments. Children who speak Spanish should be tested in both English and Spanish. The social emotional scales are completed by the child's teacher, but all other assessments listed here are conducted one-on-one in the child's school. Assessments should be scheduled to avoid meal, nap and outdoor play times. Assessment time for young children should be limited to 20 to 30 minutes per session and may increase to 30-40 minutes for older children. Training of assessors is necessary and inter-rater reliability must be a requirement before assessors are permitted to collect data independently. The following child assessments have been used in previous studies successfully. Note that each child must be assessed one-on-one with the trained assessor.

Mathematical Skills

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*

This subtest examines the child's skill in analyzing and solving practical problems in mathematics. It is an effective measure for all age levels because it progresses in difficulty and the test is ended once the child reaches too many errors in a row. Thus, the test time is dependent on ability. The median reliability coefficient alphas for all age groups for this assessment meet or exceed standards and ranged from .81 to .94 (McGrew & Woodcock, 2001). These authors also report a considerable amount of evidence for supporting the validity of the instrument.

Child Math Assessment (CMA; Starkey & Klein, 2004)

The CMA is a child assessment that measures nine key principles of early math – counting, one-set addition and subtraction, two-set addition and subtraction, geometric reasoning, construction of equivalent sets, direct measurement, shape recognition, pattern duplication, and division. It has 9 tasks presented in a hands-on form using manipulatives, which in our experience makes it very attractive to the children. There are two protocols available: Spanish and English. One of the major advantages of the CMA is that it assesses a broad range of math constructs. An extensive investigation of the CMA's psychometric properties found good reliability and validity. "Test-retest reliability over a 14-day interval is .910, and Cronbach's alpha over all tasks is .898. In addition, we administered the TEMA-3 along with the CMA in order to obtain concurrent

validity with another standardized measure of early number knowledge. We obtained significant correlations between the CMA Composite Score and the TEMA Math Ability Score (.741 -.748). This is consistent with our prediction that the CMA would correlate well, but not completely overlap, with the TEMA because the CMA assesses a broader range of informal mathematical knowledge than the TEMA." (A. Klein, personal communication, July 5, 2007)

Early Literacy

Woodcock-Johnson Tests of Achievement, 3rd Edition (Woodcock, McGrew and Mather, 2001), *Broad Reading including Subtest 1 Letter-Word Identification, Subtest 2 Reading Fluency, and Subtest 9 Passage Comprehension.* For Spanish-speakers the same battery of subtests: the *Bateria Woodcock-Munoz Pruebas de Aprovechamiento – Revisado* (Woodcock & Munoz, 1990)

This cluster of subtests provides a broad measure of reading achievement. It examines reading decoding in the Letter-word Identification subtest, reading speed and semantic processing speed in the Reading Fluency subtest, and reading comprehension in the Passage Comprehension subtest. The median reliability coefficient alphas for all age groups for this assessment meet or exceed standards and range from .81 to .94 (McGrew & Woodcock, 2001). These authors also report a considerable amount of evidence for supporting the validity of the instrument. Generally, only the Letter-word Identification subtest is used with young children in preschool and kindergarten. This test begins with matching letters, distinguishing letters from pictures, and moves to identifying letters.

Early Literacy Skills Assessment (ELSA; DeBruin-Parecki, 2005)

The ELSA is a child assessment that measures four key principles of early literacy – comprehension, phonological awareness, alphabetic principle, and concepts about print. It has 23 items presented in a children's storybook form which in our experience makes it very attractive to the children. There are two protocols that are both available in Spanish and English. One of the major advantages of the ELSA is that it assesses a broad range of language and literacy constructs including comprehension, phonological awareness, alphabetic principle, and concepts about print. An extensive investigation of the ELSA's psychometric properties conducted by an outside evaluator found good reliability and validity. "Taken in sum, these results confirm the reliability of the ELSA as a measure of children's early literacy skills. Furthermore, the consistency of the results supports the general validity of the ELSA constructs for assessing both English and Spanish-speaking populations."(p. 9, Cheadle, 2007) One potential drawback for this study is less ability to discriminate at the lower ends of the scoring which is exacerbated with younger aged children. However, the researchers found this was ameliorated in a pre, post design.

Tests of Preschool Early Literacy (TOPEL; Lonigan, Wagner, Torgesen, 2007)

This test measures abilities related to early literacy. It contains three subtests: print knowledge that measures early knowledge about written language conventions and alphabet knowledge, definitional vocabulary that measures a child's single word oral vocabulary and definitional vocabulary, and phonological awareness which examines elision and blending abilities. The scores from these subtests are combined to provide a composite score. Results are provided in raw scores, standard scores, and percentile ranks. The authors demonstrate the measure to be a valid measure of early literacy abilities. The reliability evidence for the subtests and the composite score are good and

range from .87-.96 for internal consistency, .81-.91 for test-retest, and .96-.98 for inter-scorer differences. This test is appropriate only for children age 3 through 5.

Receptive Vocabulary

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. The results of these tests are found to be strongly correlated to school success.

Social Skills

Social Skills Rating Scales - Preschool version (*SSRS*; Gresham & Elliott; 1990)

The SSRS assesses the perceived frequency and importance of children's social behaviors. The teacher rates the child by responding to the item and circling a selected response on three domains: social skills, problem behaviors, and academic competence. Reliability of the instrument is good with internal consistency for teachers on the domains ranging from .82-.95 and test re-test for teachers ranging from .84-.93. This scale has also shown content, construct, and concurrent validity.

Vineland Adaptive Behavior Scales (Sparrow, Balla, & Cicchetti, 1985)

This is an individual teacher questionnaire that assesses personal and social skills in the following domains: communication, daily living skills, socialization, and motor skills. For the classroom edition the reliability was shown as good with coefficient alpha means that ranged from .80-.95. This instrument also has shown extensive content, concurrent, and construct validity.

Child Assessment Training

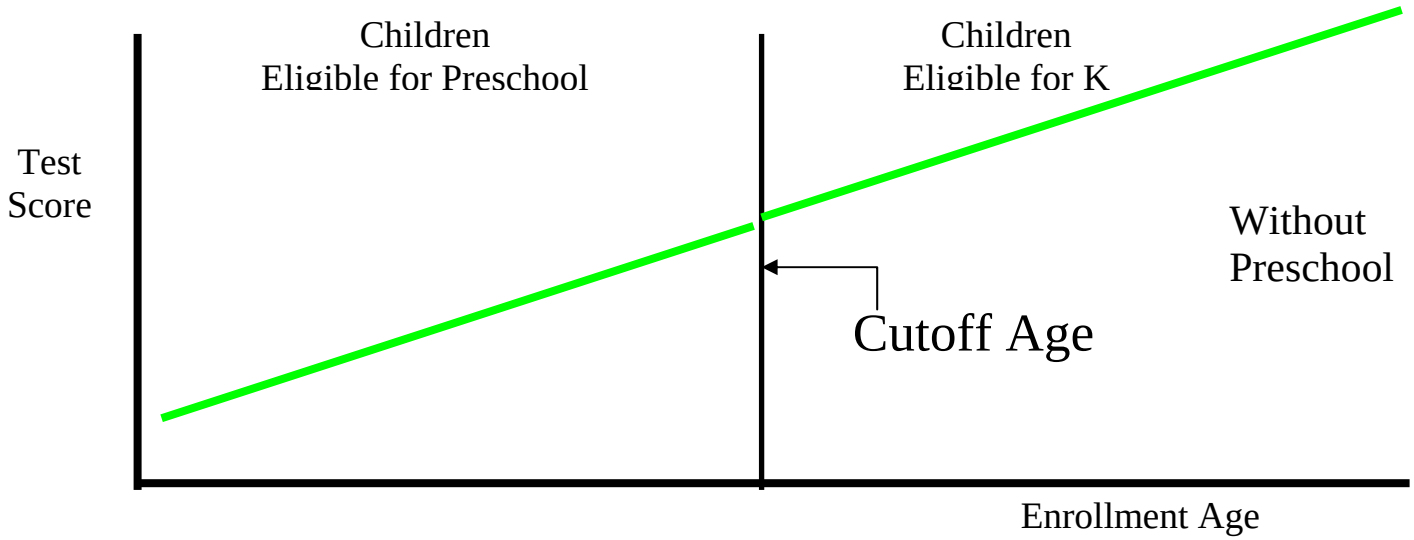
Assessors are trained on each child assessment, usually over the course of 2 days, and then shadow score in practice assessments until they reach 100% agreement with the trainer. Training includes issues related to assessing children in school environments, confidentiality, protocol for reporting instances of child abuse, and professional etiquette as well as training specific to the assessment instruments and sampling procedures. Refresher training occurs again just prior to each round of child assessments.

Appendix C

Regression-Discontinuity Design Representation

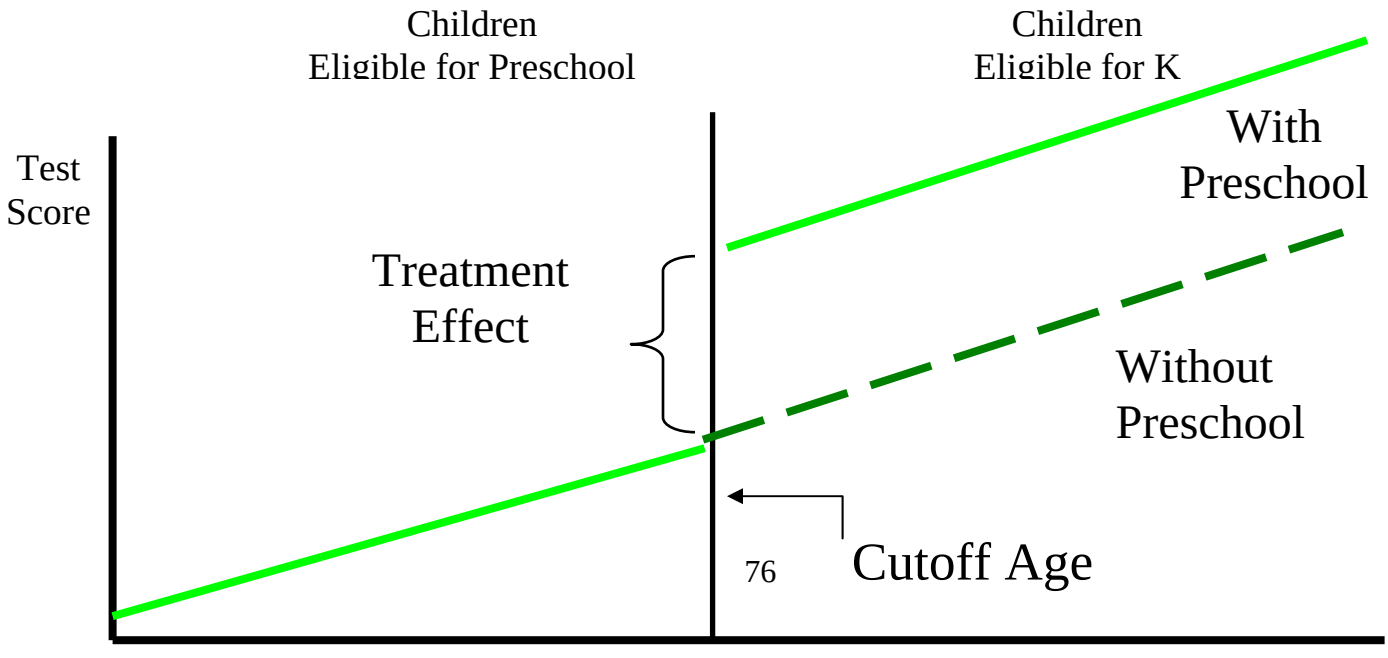
Test Scores by Age

This graph represents the expected performance of children on test scores when the only difference at the cutoff line is age.



The Effect of Preschool

This graph demonstrates the discontinuity at the cutoff age on test scores when the treatment of preschool is provided.



References

Ayers, S., Stevenson-Boyd, J., & Frede, E. (2007). *The Early Learning Scale.* New Brunswick, NJ: National Institute for Early Education Research.

Barnett, W. S. (1993). *Lives in the balance: Age 27 benefit-cost analysis of the High/Scope Perry preschool study*. Ypsilanti, MI: High Scope Press.

Barnett, W. S., Frede, E., Cox, J. O., & Black, T. (1994). Using cost analysis to improve early childhood programs. In W. S. Barnett (Ed.), *Cost analysis for education decisions: Methods and examples* (pp. 145-182). Greenwich, CT: JAI Press.

Barnett, W. S., Frede, E. C., Mobasher, H., & Mohr, P. (1987). The efficacy of public preschool programs and the relationship of program quality to efficacy. *Educational Evaluation and Policy Analysis, 10*(1), 37-49.

Barnett, W. S., Howes, C., & Jung, K. (2008). *California's preschool program: Quality and effects of on children's learning.* (Working Paper). New Brunswick, NJ: NIEER.

Barnett, W. S., Jung, K., Wong, V., Cook, T. & Lamy, C. (2008). Effects of five preschool programs on early learning. (Working Paper). New Brunswick, NJ: NIEER.

Barnett, W. S., & Masse, L. N. (2007). Comparative benefit-Cost Analysis of the Abecedarian Program and its policy implications. *Economics of Education Review*, *26*, 113-25.

Belfield, C. R. (2005). *An economic analysis of preschool in Louisiana.* Washington, DC: Pre[K]Now.

Belfield, C. R. (2006a). *An economic analysis of four-year-old kindergarten in Wisconsin: Returns to the education system.* Washington, DC: Pre[K]Now.

Belfield, C. R. (2006b). *An economic analysis of preschool in Arkansas.* Washington, DC: Pre[K]Now.

Belfield, C. R. (2006c). *Does it pay to invest in preschool for all? Analyzing return-on-investment in three states.* (NIEER Working Paper). New Brunswick, NJ: NIEER.

Belfield, C. R., Nores, M., Barnett, W.S. & Schweinhart, L.J. (2006) The High/Scope Perry Preschool program: Cost-benefit analysis using data from the age-40 follow-up. *Journal of Human Resources, 41*(1), 162-90.

Brown, G., Scott-Little, C., Amwak, L., & Wynn, L. (2007). A review of methods and instruments used in state and local school readiness evaluations. Available: http://ies.ed.gov/ncee/edlabs/projects/project.asp?id=64

Cheadle, J. (2007). Early Literacy Skills Assessment psychometric report: For both English and Spanish Versions. Ypsilanti, MI: High/Scope Press. Available: http://www.highscope.org/file/Assessment/ELSAJacobs.pdf

DeBruin-Parecki, A. (2005). *Early Literacy Skills Assessment.* Ypsilanti, MI: High/Scope Press.

Dichtelmiller, M. L., Jablon, J. R., Dorfman, A. B., Marsden, D. B., & Meisels, S. J. (2000). *Work sampling in the classroom: A teacher's manual.* Ann Arbor, MI: Rebus Inc.

Dunn, L. & Dunn, L. (1997). *Peabody Picture Vocabulary Test, Third edition (PPVT- 111).* Circle Pines, MN: American Guidance Service.

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

Epstein, A. S., Schweinhart, L. J., DeBruin-Parecki, A., & Robin, K. B. (2004). *Preschool assessment: A guide to developing a balanced approach.* New Brunswick, NJ: NIEER.

Feuer, M., Towne, L. & Shavelson, R. J. (2002). Scientific culture and educational research. *Educational Researcher, 31(8),* 4-14.

Frede, E. (2005). *Assessment in a continuous improvement cycle: New Jersey's Abbott Preschool Program.* Paper prepared for the National Early Childhood Accountable Task Force.

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).*

Frede, E., Barnett, W. S., Jung, K., Lamy, C. E., & Figueras, A. (2007). *Abbot preschool program longitudinal effects study (APPLES): Interim report.* Retrieved September 10, 2008 from http://nieer.org/resources/research/APPLES.pdf.

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 outcomes.* Paper presented at the annual National Association for the Education of Young Children conference, Washington, DC.

Frede, E., Weber, M., Hornbeck, A., Stevenson-Boyd, J., & Colón, A. (2005). *Preschool ClassroomMathematic Inventory (PCMI)*. New Brunswick, NJ: National Institute for Early Education Research.

Gilliam, W. S. & Zigler, E. F. (2000). A critical meta-analysis of all evaluations of state-funded preschool from 1977-1998: Implications for policy, service delivery and program evaluation. *Early Childhood Research Quarterly, 15(4),* 441-473.

Gilliam, W. S. & Zigler, E. F. (2004). *State efforts to evaluate the effects of prekindergarten: 1977-2003.* Retrieved September 10, 2008 from http://nieer.org/resources/research/StateEfforts.pdf.

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

Greshman, F., & Elliot, S. (1990). *Social Skills Rating System*. Bloomington, MN: Pearson Assessments.

Harms, T., Clifford, R., & Cryer, D. (2005). *Early Childhood Environment Rating Scale (ECERS-R),* revised edition. New York, NY: Teacher College Press.

Karoly, L. A., Kilburn, K., & Cannon, J. S. (2005). Early childhood interventions: Proven results, future promise. Santa Monica, CA: The RAND Corporation.

Klein, A., & Starkey, P. (2000). *Child Math Assessment.* University of California, Berkeley.

LaParo, K. M., Pianta, R. C., & Stuhlman, M. (2004). The classroom assessment scoring system: Findings from the preschool year. *Elementary School Journal, 104(5),* 409-426.

Levin, H. M., & McEwan, P. J. (2001). *Cost-effectiveness analysis* (2nd ed.). Thousand Oaks, CA: Sage.

Lonigan, C. J., Wagner, R. K., & Torgesen, J. K. (2007). *Test of preschool early literacy: TOPEL.* Austin, TX: Pro-ed.

Lynch, R. (2004). *Exceptional returns: Economic, fiscal, and social benefits of investment in early childhood development.* Washington, DC: Economic Policy Institute.

McGrew, K. S. & Woodcock, R. W. (2001). *Woodcock-Johnson III Technical Manual*. Itasca, IL: Riverside Publishing.

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 .

National Association for the Education of Young Children (NAEYC) and National Association of Early Childhood Specialists in State Departments of Education (NAECS/SDE) (2003). *Early childhood curriculum, child assessment and program evaluation: Building an accountable and effective system for children birth through age eight. A joint position statement of NAEYC and NAECS/SDE.* Washington, DC: NAEYC.

Pianta, R., Howes, C., Burchinal, M., Bryant, D., Clifford, R. M., Early, D. M., & Barbarin, O. (2005). Features of preschool programs, classrooms, and teachers: Prediction of observed classroom quality and teacher-child interactions. *Applied Developmental Science, 9 (3),* 144-159.

Pianta, R. C., La Paro, K. M., & Hamre, B. (2006). *Classroom Assessment Scoring System (CLASS)*. Charlottesville, VA: Center for Advanced Study of Teaching and Learning.

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.

Ritchie, S., Howes, C., Kraft-Sayre, M., Weiser, B. (2002). *Snapshot*. Los Angeles: University of California, Los Angeles.

Schultz, T. & Kagan, S. L. (2007). *Taking stock: Assessing and improving early childhood learning and program quality. The report of the National Early childhood Accountability Task Force.* Available: http://www.pewtrusts.org/our_work.aspx?category=102

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. Available at http://www.highscope.org/Research/PerryProject/perrymain.htm

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* New York, NY: Houghton Mifflin Company.

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.

Sparrow, S. S., Balla, D. A., & Cicchetti, D. V. (1985). *Vineland Adaptive Behavior Scales.* Circle Pines, MN: American Guidance Service.

Strategic Research Group (2008). *Assessing the effectiveness of Tennessee's pre-kindergarten program: Second interim report.* Retrieved September 10, 2008 from http://www.comptroller1.state.tn.us/repository/RE/prekeval08.pdf.

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,* 126-44.

Wong, V. C., Cook, T. D., Barnett, S. W., & Jung, K. (2007) An effectiveness-based evaluation of five preschool programs. *Journal of Policy Analysis and Management,* 27(1), 122-154.

Woodcock, R. & Munoz, A. F. (1990). *Bateria Woodcock-Munoz: Pruebas de habilidad cognitive-Revisada*. Chicago, IL: Riverside Publishing.

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