Technical Report for the City of Seattle.
A Review of the Evidence on Preschool Programs and a Comparison of Selected State and City Programs

Prepared by
Milagros Nores,
Steve Barnett,
Allison Friedman-Krauss
and Jessica Francis[1]

Introduction

To inform the evaluation of Seattle Pre-K we reviewed research on the effectiveness of preschool education, with particular attention to studies of large-scale public preschool programs administered in states and cities in the United States that most closely resemble the Seattle model. We describe these programs and their effects in detail. All of the programs have some of the characteristics of "high-quality" programs, and all have been found to have some positive effects on children's learning and development. Positive effects of large-scale public programs have been found for language, literacy, math, executive function, and behavior, as well as reduced school failure as indicated by grade repetition and special education placements. In general, more educationally intensive programs (which does not mean overly academic and didactic) have larger and more persistent impacts. Weaker programs have produced weaker effects that are, at best, of very small size in the long run.

In response to the request of the City of Seattle, we focused greater attention on studies of programs in Georgia, Oklahoma, and New Jersey, and in the cities of Boston, Denver, and San Antonio. We also draw attention to randomized trials of programs in Tennessee and Providence, Rhode Island. All of these programs have elements of high quality, but they vary considerably along many dimensions. The studies also vary in terms of their design, from highly rigorous to much less rigorous approaches, depending on local constraints. They are quite similar with respect to measures of both classroom quality and children's learning and development.

We present options and derive recommendations for the design and overall parameters of the study, and for specific details regarding instruments to assess quality and educational effectiveness.

[1] Authors are extremely grateful for the assistance provided by Griffin Wood.

1. Review of the evidence on high quality existing programs

Studies in the United States over many decades have found that high quality early childhood education can produce large and persistent benefits for children's learning and development. These benefits in turn yield important improvements in children's later life outcomes, including school progress and educational attainment, employment and earnings, socialization and social behavior (including reduced delinquency and crime), and even improved health.

The results of large numbers of studies of preschool education have been summarized by meta-analyses, which are basically statistical summaries of findings. These statistical summaries take into account the strengths and weaknesses of studies when summarizing them, giving more weight to stronger studies or separately summarizing the effects of strong and weak studies. In order to summarize results across studies using different measures, all of the findings are converted to a common measure, or effect size, which is relative to how much variation there is in a given domain (in technical terms this is the standard deviation).

We caution that the meta-analysis approach of computing uniform effect sizes does not really mean that effects on language or math or various aspects of literacy can truly be compared to each other, much less that they can be perfectly compared to effects on classroom behavior. Nevertheless, as a first approximation for pooling results across studies, it is useful. We report standardized effect sizes both from meta-analyses and when we discuss individual studies to facilitate cross study comparisons. As a practical matter, it helps to understand that the gap in achievement at various points during the school career between children in poverty and others is an effect size of between 0.50 and 1.00. Therefore, effect sizes in this range can be considered quite large. Effect sizes of 0.25 (if persistent) are substantial.

Multiple meta-analyses find average effect sizes for initial preschool program impacts of about 0.50 on a wide range of outcome measures, including achievement and other cognitive tests (Camilli, et al., 2010; Gorey, 2001; Guralnick & Bennett, 1987; McKey, Condelli, Ganson, Barrett, McConkey & Planz, 1985; Nelson , Westhues & MacLeod, 2003; Ramey, Bryant & Suarez, 1985; White & Casto, 1985). Of course, this is not the average from public programs that provide one year of preschool education, but across a wide range of programs, including many that provide two or more years of early education. Although effects on cognitive tests tend to decline after children enter school, they do not (on average) disappear (Camilli et al., 2010).

Average effects on social and emotional development tend to be smaller, perhaps because fewer programs focus on such outcomes (Camilli et al., 2010; McKey et al., 1985; White & Casto, 1985). Randomized trials have found that alternative approaches to preschool education can have similar effects on cognitive development while having very different effects on social and emotional development (Scwheinhart & Weikart, 1997; Barnett et al., 2007).

Beyond providing evidence of the large impacts of early childhood programs, these studies have found that effects vary with program quality, quantity, content, range of services, and type of instruction (e.g. Anderson et al., 2003, Camilli, et al.. 2010; Gorey, 2001; Karoly, Kilburn and Cannon, 2005; Nelson et al., 2003; NICHD Early Childcare Research Network Study, 2002). Camilli and colleagues (2010) found that programs that emphasized intentional teaching, small group time, and individualization produced larger effects. More broadly, public programs have smaller effects than more intensive model programs, while state-funded programs tend to outperform Head Start (Kay, & Pennucci, 2014). Meta-analyses also find

that research design has little impact on average effect size, but that the choice of outcome measures has a large impact on results--larger effects are found on narrower measures of discrete skills that are more easily mastered in a short time (Shager, Schindler, Magnuson, Duncan, Yoshikawa, & Hart, 2012).

High Quality Landmark Studies: the Big Three

Statistical summaries of the research are useful, but limited in the extent to which they permit detailed examination of all of the elements that contribute to differences in findings regarding program effectiveness. Moreover, relatively few studies among the hundreds conducted combine highly rigorous designs with long-term follow-up. Three such studies stand out in the literature. Two are randomized trials of very intensive small-scale programs that produced large effects. The third is a quasi-experimental study of a large-scale public program operated by the Chicago Public Schools. This last program is the most similar to strong programs currently offered on a large scale by states and cities (e.g., Gormley et al., 2008; Frede & Barnett, 2011). All three programs served only low-income populations. We review each of the landmark studies in detail below.

The HighScope Perry Preschool Program study is quite likely the most widely cited and influential preschool study ever. Its influence can be explained by the simplicity of its design which makes it easy to understand, its remarkable findings, and its very long-term follow-up of almost all the original participants. The study began in 1962 by offering a program consisting of half-day classes five days a week, and weekly home visits to highly disadvantaged children in one small neighborhood. Home visits focused on one-on-one tutoring. The teachers had bachelor's or higher degrees, and the teacher-child ratios were 1 to 6 or 7. It used a participatory education model with a balance of child-initiated and teacher-initiated activities. The study's 123 participants were randomly assigned at age 3 (with a small number in the first year at age 4) to either the program or a control group (which in the early 1960s had no center-based alternatives). The local public school operated the program, and the vast majority of the program group attended for two school years.

To date, children have been followed through age 40 (Schweinhart et al., 2005). Strong short-term effects were found on cognitive development. In the early school years, effects were found on achievement and classroom behavior. Increased achievement was found throughout the elementary and secondary years. Decreased delinquency, crime, and risky behaviors (smoking and teen pregnancy) were found in adolescence. Positive effects for adults included higher secondary school completion rates, higher employment rates, and significantly higher median annual earnings. Benefit-cost studies of the Perry Preschool have taken into account educational, crime, welfare, and labor market outcomes. These outcomes result in an estimated economic return between 6 and 15 dollars per dollar invested, with much of the return due to reductions in crime (Belfield, Nores, Barnett and Schweinhart, 2006; Heckman, Moon, Pinto, Savelyev & Yavitz, 2010).

The Abecedarian Study began in the 1970s, with slightly more than 100 disadvantaged children randomly assigned to a full-day educational child care program or a control group at birth, and has followed those participants through age 30 (Campbell et al., 2012). As with the Perry study, the design is easily understood and the vast majority of the participants remained in the study over time. Children in the program received intensive high-quality center-based child care eight hours a day through the age of five. The program used child-centered, individualized approaches, and emphasized development of cognition and language skills. As with the Perry Preschool program, there was strong support for high quality program implementation by a dedicated team who worked with the teachers routinely, and monitored classroom practices and results. For the Abecedarian program, mid-course adjustments were

made to the curriculum, based on reports that program children had higher rates of problem behaviors in the early years (and this problem was alleviated). In this study, children could have attended another center-based child care program, and some evidence indicates that this reduces the advantage of the program children over the control condition children.

Strong results were evident immediately and throughout the follow-up. In the short term, children in the program scored higher on IQ and achievement tests, and they continued to score higher in reading and math through age 21 (Campbell, et al. 2001). In addition, rates of special education and grade repetition were much lower, and later educational attainment was higher. Abecedarian program participants were also significantly more likely to hold a skilled job at age 21, and less likely to smoke or use drugs. Effects on health behaviors and health were found at age 30 (Muennig et al., 2011). Benefit-cost analyses for the Abecedarian study found an estimated return of about 4 dollars per dollar invested (Barnett & Masse, 2007). No effects on crime were found for the Abecedarian program, which together with its much higher cost led to a low return in comparison to the other studies discussed here.

The Chicago Child-Parent Center (CPC) study began in the 1980s. This study used matched neighborhoods and statistical controls to estimate program effects of a part-day preschool program at ages 3 and 4 (Reynolds, Temple & Ou, 2010). The study compared about 1,000 children who completed preschool and kindergarten in the Child-Parent Centers (CPC) operated by Chicago Public Schools to about 500 children in similar neighborhoods where the preschool program was not offered. These two groups of children were from comparably low socioeconomic backgrounds. The preschool program had a ratio of 17:2; required teachers with BA degrees with early childhood certification; and included regular staff development activities, health and social services, and free meals. It used a curriculum defined around a structured set of educational activities emphasizing reading and math skills, and included parent participation opportunities and parent support. Note that because the comparison group is not identified until kindergarten entry, there is no pre-test of initial abilities that can be used to guarantee that the preschool and comparison groups had the same abilities at age 3 or to adjust for any differences that there might have been from the start.

Results of the CPC program have been remarkably similar to those of the more intensive Perry Preschool program in that effects are found for all of the same outcomes--cognitive, academic, and behavioral--but the effect sizes tend to be smaller. This suggests a dose-response relationship that is not evident in the meta-analyses. Short-term results for the CPC study (Reynolds, 1995) include increased cognitive readiness at kindergarten entry (a difference of about three months), and better reading and math test scores. These differences in test scores continue throughout school, and there are reductions in retention and special education placement as well. In addition, two-years of preschool seemed to improve kindergarten readiness an equivalent of two months. The added effects of two years are modest during the school years, but appear to be larger for long-term adult outcomes. Longer-term effects (Reynolds, 2000; Reynolds, et al., 2001 & 2002) include lower rates of juvenile delinquency and arrests, and increased rates of high school completion. Later follow-ups find increased college attendance; lower incarceration, arrests and conviction rates; and increased earnings (Reynolds, et al., 2007).

The benefit-cost ratio is estimated to be about 10 to 1, similar to that for the Perry Preschool (Reynolds et al., 2011). Although benefits are lower, the cost is proportionately lower, as well. Cost of the CPC program is similar to that of today's public preschool programs that operate with pay scales comparable to that of the public schools, well-qualified teachers, and class sizes of around 20. As we discuss below, similar large-scale public programs operating today produce initial effects on cognitive abilities that are

comparable in size to the effects of the Chicago program (Barnett, 2011). Because initial effect size is a good predictor of long-term effects, this provides some confidence that current programs can produce similar long-term outcomes, as well. It also indicates that a quasi-experimental design can yield results very comparable to those of a randomized trial, albeit with a much larger sample size. Nevertheless, the CPC study has not been as influential as the Perry and Abecedarian studies, despite its greater relevance to current public programs, because its design inspires less confidence in the findings.

Large-Scale Public Programs

Research finds that state and local public pre-K programs, almost without exception, improve readiness for school and produce long-term effects. However, effect sizes vary not just by type of outcome measure, but also by type of program. High-quality, well-defined education programs provided by public schools have been found to produce the largest effects on child development among typical large-scale programs for 3- and 4-year-olds (Consortium for Longitudinal Studies, 1983; Deutsch, Deutsch, Jordan, & Grallow, 1983; Deutsch, Taleporos, & Victor, 1974; Frede, 1998; Jordan, Grallo, Deutsch, & Deutsch, 1985; Schweinhart, Montie, Xiang, Barnett, Belfield, & Nores, 2005; Kay & Pennucci, 2014). Whether the public school advantage is due to greater infrastructure and support; higher compensation levels and better working conditions; increased continuity with the later grades; or some other factor or combination of factors, cannot be determined from the studies available. It does seem that programs do not need to be directly operated by the public schools to produce the same effects. Also, public school auspices do not guarantee large effects.

One of the most rigorous evaluations of large scale public preschool programs conducted to date is the national impact study of Head Start (Puma et al., 2010; Puma et al., 2012). This evaluation employed a randomized trial with 4,667 children with a nationally representative sample of programs. Initial effects were small to moderate, about .15 for language and math (though effects on math were not statistically significant) and about .30 for literacy. As can be seen in Table 1a these cognitive effects are considerably smaller than in most studies of state and local programs. Social behavior effects of about .15 were found for 3-year-olds but not 4-year-olds. By end of the following year (kindergarten for those who were 4) the effects were essentially zero. No consistently better results were found for any subgroup of children or programs. These results also are relatively disappointing compared to those of most state and local programs as discussed below.

The most recent studies of large scale, state-funded and city-funded pre-K programs provide estimates of short-term effects overall and for subgroups that are consistent with past studies. Effect sizes for initial impacts range from about 1.0 sd for narrow, basic literacy skills to about .25 to .35 sd for broader measures of language and mathematics (Barnett, Howes, & Jung, 2008; Gormley, Gayer, Phillips, & Dawson, 2005; Gormley et al., 2008; Hustedt, Barnett, Jung, & Figueras, 2008; Hustedt, Barnett, Jung, & Thomas, 2007; Lipsey, Farran, Hofer, Bilbrey & Dong, 2011; Peisner-Feinberg & Schaaf, 2011; Wong, Cook, Barnett & Jung, 2008). Significant positive effects are present for boys and girls, for all ethnicities, and for children from families at all income levels. Effects tend to be larger for children from lower-income backgrounds and from homes where English is not the first language.

State and City pre-K evaluations. Summary and selected highlights.

Table 1a below summarizes findings regarding initial effects of State and City pre-K programs. Overall, the findings across evaluations for this set of cities and states are quite consistent. The vast majority of estimated effects are positive, and this also holds for the much smaller number of estimates of effects

on executive function (which is not shown in Table 1a). The estimated effects do vary considerably across sites. In general, the regression discontinuity design tends to generate larger effect size estimates than other research designs, and the more commonly used measures of literacy that focus on such basic discrete skills as letter recognition produce larger estimated effects. The largest effect size estimates for the most difficult outcome, which is broad language development, are produced by programs that are mature and the most strongly resourced: Boston, New Jersey's Abbott program, and Oklahoma's universal pre-K program. It is interesting that none of these programs is tightly focused on only disadvantaged children.

The average effect size across the evaluations in Table 1a is about 0.30 for math, 0.20 for language and 0.60 for literacy. Another way to look at the effect sizes in Table 1 is to see how they are distributed by size. As noted earlier, the literacy effect sizes are more often very large. How often are estimated effects large enough to generate confidence that effects will persist and that they can be detected in evaluations with reasonable sample sizes? For both math and literacy, effect sizes are mostly above 0.25 and rarely below 0.10. For language, about half the effect sizes are above 0.25, but it also is common for them to be below 0.10. Therefore, it is important that evaluations measure both language and literacy, and in planning it should be recognized that small sample size may be especially constraining for detecting effects on language. However, we note that effect sizes of less than .10 for language are likely to be associated with a lack of meaningful long-term outcomes in later achievement and school progress, as in the national Head Start impact evaluation. We also note that the most commonly used literacy measures (such as the Woodcock Johnson) produce larger estimates than less commonly used measures. As we discuss later, it seems wise for evaluations to use a combination of measures that assess skills that are both easy and difficult to affect with preschool education.

Evaluation results for follow-up effects after entry to kindergarten (and beyond) are summarized in Table 1b. Clearly, effects are smaller in follow-up. However, effects also vary across programs from essentially zero (both times in randomized trials) to around .10 and .20. Effect sizes do not vary by type of measure (math or language and literacy) at older ages. Programs found to have the largest initial impacts (Table 1a) tended to have the largest long-term effects in Table 1b, as well. As we discuss in Section 2 of the literature review, evaluation design may contribute to differences in estimated effects, as well.

Table 1a. State and local pre-K effect sizes at end of pre-K or by Kindergarten, by state/city and study.

Location	Year	Learning outcome			Reference
		Math	Language	Literacy	
State					
Arkansas	K	.27	.28	1.00	Barnett, et al.(2013)
California	K	.34	.39	1.19	Barnett, et al.(2013)
Florida(*)	K	.25	.23		King, Cappellini, & Rohanic (1995)
Georgia	end of pre-k	.18	.06	.14	Peisner-Feinberg, Schaaf, & LaForett (2013)
Georgia	K	.69	.01	.93	Peisner-Feinberg, Schaaf, LaForett, Hildebrandt, & Sideris (2014)
Michigan	K	.47	-.13	.96	Wong et al. (2008)
Michigan	K	.93	.19	1.52	Barnett et al. (2013)
Michigan	K	.51	.45		Florian, Schweinhart, & Epstein (1997)
North Carolina	end of pre-k	.30	.19	.21	Peisner-Feinberg & Shaaf, 2008
North Carolina	K	.07	.27	.93	Peisner-Feinberg & Schaaf (2011)
New Jersey	K	.23	.36	.50	Wong et al. (2008)
New Jersey	K	.13-one year .29-two years	.22-one year .41-two years	.11-one year .14-two years	Frede, Jung, Barnett, & Figueras (2009)
New Mexico	K	.33	.18	1.06	Barnett et al. (2013)
Oklahoma	K	.43	.31	.57	Barnett et al. (2013)/Wong et al. (2008)
South Carolina	K		.05	.79	Barnett et al. (2013)/Wong et al. (2008)
Tennessee	end of pre-k	.32	.31	.46	Lipsey, Hofer, Dong, Faran, & Bilbrey (2013)
Texas	K	.05			Kuhne (2008)
West Virginia	K	.10	.16	.82	Barnett et al. (2013)/Wong et al. (2008)

National representative sample	K	.40	.73		Magnuson, Ruhm, & Waldfogel (2004)
National representative sample[a]	K	.15		.18	Magnuson, Ruhm, & Waldfogel (2007b)
Cities					
Providence, RI	K	.24	.05	.27	Barnett, Francis, & Youn (unpublished)
Boston	K	.59	.44	.62	Weiland & Yoshikawa (2013)
Chicago	K	.39	.44		Reynolds, Temple, & Ou (2010)
Denver	K	.10		.11	Robertson, McClelland, Palaich, Rooney, & Workman (2015)
Tulsa, OK	K	.36		.99	Gormley et al. (2008)
Head Start:					
Head Start[b]	end of 3-year-old pre-k	.18	.17	.35	Ludwig & Phillips (2008)
Head Start[c]	end of pre-k	.15	.08	.32	Ludwig & Phillips (2008)
Head Start[b]	spring of K	-.04	.01	.01	Puma et al. (2010)
Head Start[c]	spring of K	.01	.04	-.01	Puma et al. (2010)
Head Start[d]	K		.32	.33	Abbott-Shim et al. (2003)
Head Start	K	-.07		-.14	Loeb et al. (2007)
Head Start	K	.01		-.03	Magnuson, Ruhm, & Waldfogel (2007a)
Head Start[e]	K	.02		.08	Magnuson, Ruhm, & Waldfogel (2007b)

a. Public school children only. b. 3-year-old cohort. c. 4-year-old cohort. d. Only one year of HS. e. Low income children only.
Source: Adapted and updated from Barnett, 2013. There are no robust evaluations available for San Antonio.

Table 1b. State and local pre-K effect sizes beyond the end of pre-K, by state/city and study.

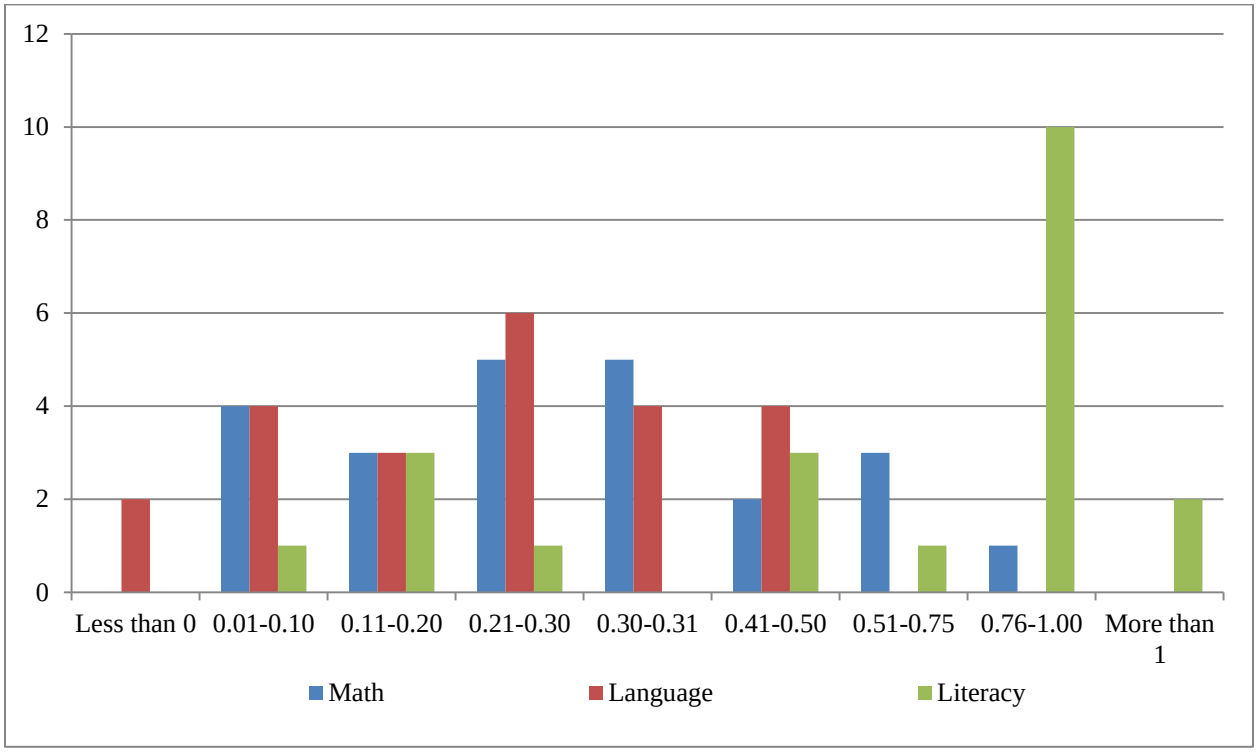
Location	Year	Learning outcome		Reference
		Math	Language/Literacy	
States				
Tennessee	end of K	.02	-.09	Lipsey, Hofer, Dong, Faran, & Bilbrey (2013)
Georgia	4th grade	.19	.19	Bartik (2013)
Georgia	1st grade		.24	Pilcher & Kaufman-McMurrain (1996)
Georgia	4th grade	.03	.03	Fitzpatrick (2008)
Georgia/Oklahoma	8th grade	.06		Cascio &Schanzenbach (2013)
Michigan	7th grade	0	0	Maloffeva, Daniel-Echols, & Xiang (2007)
New Jersey	1st grade	.18-one year .26-two years	.18-one year .38-two years	Frede, Jung, Barnett, & Figueras (2009)
New Jersey	2nd grade	.24-one year .44-two years	.19-one year .30-two years	Frede, Jung, Barnett, & Figueras (2009)
New Jersey	4th grade	.17-one year .37-two years	.12-one year .26-two years	Barnett, Jung, Youn, & Frede, (2013)
New Jersey	5th grade	.14-one year .29-two years	.18$_{\text{-one year}}$ .22$_{\text{-two years}}$	Barnett, Jung, Youn, & Frede, (2013)
North Carolina	3rd grade	.24	.17	Ladd, Muschkin, & Dodge (2014)
North Carolina	3rd grade	.12	.14	Peisner-Feinber & Shaaf (2010)
Tennessee	1st grade	-.05	-.01	Lipsey, Hofer, Dong, Faran, & Bilbrey, (2013)
Tennessee	2nd grade	-.14	-.12	Lipsey 2014
Texas	3rd grade	.05	.05	Huston, Gupta, & Schexnayder (2012)
Texas	3rd grade	.09	.08	Texas Education Agency (1995)
Texas	3rd grade	.04-one year .10-two years	.04-one year .09-two years	Kuhne (2008)
Texas	4th grade	.04	.05	Kuhne (2008)

Texas	5th grade	.03	.04	Kuhne (2008)
Texas	6th grade	.01	.02	Kuhne (2008)
Texas	7th grade	.01	.02	Kuhne (2008)
Texas	8th grade	.01	.02	Kuhne (2008)
Washington	3rd grade	.14	.17	WSIPP (2014)
Washington	4th grade	.16	.26	WSIPP (2014)
Washington	5th grade	.16	.23	WSIPP (2014)
National representative sample	1st grade	.05	.03	Magnuson, Ruhm, & Waldfogel (2007a)
National representative sample	3rd grade	.07	.06	Magnuson, Ruhm, & Waldfogel (2007a)
National representative sample	1st grade	.03	.03	Magnuson, Ruhm, & Waldfogel (2007b)
Cities				
Chicago	3rd-6th grade	.25	.26	Reynolds (2000)
Chicago	8th grade	.17	.18	Reynolds (2000)
Chicago	8th grade	.29	.33	Reynolds (1997)
Tulsa, OK	3rd grade	.18	.09	Hill, Gormley, & Adelstein, (2012)
Head Start				Head Start:
Head Start[b]	1st grade	-.01	.03	Magnuson, Ruhm, & Waldfogel (2007b)
Head Start[c]	1st grade	.08	.05	Puma et al. (2010)
Head Start[d]	1st grade	.04	.06	Puma et al. (2010)
Head Start[c]	3rd grade	.01	.04	Puma et al. (2012)
Head Start[d]	3rd grade	-.02	.09	Puma et al. (2012)
Head Start	7th-12th grade	-.03		Aughinbaugh (2001)

a. Effect size per $100 increase in expenditure. b. Low income children only. c. 3-year-old cohort. d. 4-year-old cohort.

Source: Adapted and updated from Barnett, 2013.

Figure 1. Distribution of effect sizes by domain measured.



2. Comparison of program features

Pre-K programs across the United States vary along a multitude of dimensions. These include funding (formulas, sources, amounts), oversight and supervision, standards and policies, and, of course, actual quality of the services delivered to children. The consistency of the findings reported in the previous section is somewhat surprising when one takes into account that degree of variation. However, this is consistent with the failure of meta-analyses and other studies to find strong associations between outcomes and various program and policy features. This suggests to us that it is an overall constellation of features that is associated with differences in outcomes, rather than any individual program or policy feature. Of course, very low funding levels (as in South Carolina, for example) may be expected to preclude having a strong set of program features overall.

The City of Seattle has expressed a particular interest in the outcomes and features of programs in Georgia, New Jersey, and Oklahoma, and in the city-level programs of Boston, Denver, and San Antonio. We first provide a brief overview of these programs and their evaluations and program outcomes. We then provide detailed information for each of the programs regarding program characteristics and key features of their policies and practices.

2.1 Overview of State Programs

Georgia's Bright from the Start Pre-K Program. Georgia's pre-K program is lottery funded and, in theory, available to all 4-year-olds in the state. However, as the program has grown, lottery funds proved insufficient to serve all children who wish to enter (at least in some locations). The program was established in 1993 and by 2013 it enrolled more than 80,000 children (60 percent of 4-year olds) in 6.5 hour per day, 5 day a week program operating over the school year (Barnett, Carolan, Squires, Clarke Brown, & Horowitz, 2015). Georgia meets 8 out of the 10 quality benchmarks for policies that are a focus of the State of Preschool Yearbook for 2014, which places it among the top half of state programs with respect to policies.[2] The program uses a mixed delivery system that includes public schools, private providers, and Head start. In an assessment of classroom quality, only 2 percent scored 'good' or better on the ECERS-R (average score 3.6) while 7 percent scored 4 or better on instructional support as measured by the CLASS (average score was 2.8). Another concern we note is that there is substantial turnover among the children served during a year, as the program served more than 94,000 children at some point during the same year it had 80,000 places available.

The Georgia program has had multiple evaluations from its earliest years (Raden, 1999). The first study compared children who attended to a matched group who had not attended at the end of Kindergarten, using a relatively small sample (n= 517). It found effects on multiple domains of development and fewer absences and grade retentions for those who had attended. Early process evaluations indicated concerns with uneven quality and uneven parent satisfaction (associated with these differences) even though, overall, there was high reported parent satisfaction. A much larger longitudinal evaluation was conducted from 1996-2001, but it did not have a no-program comparison group, and it found no consistent predictors of differences in child outcomes among the program features that varied. More recent studies include longitudinal and regression discontinuity studies (comparing children who just miss the age cut-off for entry with those who just make it, which can only provide an estimate of impacts at kindergarten entry, not over time). The longitudinal evaluation found that children made substantial gains that were, in many domains, greater than might have been expected based on national norms for the assessments, but there was no comparison group. An evaluation of the program has shown that children in the program evidenced significant growth during their pre-k year in language and literacy skills, math skills, general knowledge, and behavioral skills (Maxwell, Early, Bryant, Kraus, Hume & Crawford, 2009). The positive effects of the program were present for children from different income and language backgrounds.

The regression discontinuity design (RDD) arguably is the strongest design to be applied in Georgia in terms of providing a sound comparison. In terms of Kindergarten readiness, the evaluation found significant positive effects on children's literacy skills across most areas (letter knowledge, letter-word identification, phonological awareness, and phonemic awareness), on math problem-solving and counting, and on general knowledge. No significant effects were found on children's language (WJ vocabulary), social skills or problem behaviors (Peisner-Feinberg, Schaaf, LaForett, Hildebrandt & Sideris, 2014).

[2] The State of Preschool Yearbook is a yearly score card on the States progress on access, quality, and funding for preschool education for 3- and 4-year-olds. It includes the following 10 quality benchmarks: a) the presence of comprehensive early learning standards, b) whether lead teachers are required to have a BA, c) whether lead teachers are also required to have specialized training (e.g. a CDA), d) whether assistant teachers are required to have a CDA, e) whether teachers are provided in-service training of a at least 15 hours per year, f) whether maximum class sizes of 20 or lower are mandated, g) whether maximum child/teacher ratios of 10 to 1 or lower are mandated, h) whether screening is provided for vision, hearing, health, and at least 1 support service, i) whether at least 1 meal is provided per day, and j) whether site visits occur as part of the monitoring strategy.

New Jersey's Abbott Preschool Program. This program has especially high standards and targets 3- and 4-year-olds in the 31 poorest school districts in the state, known as the Abbott districts.[3] Although these districts are considered high-poverty for New Jersey, the New Jersey standard probably would encompass most urban school districts in the nation, as the standard is 40 percent or more of children qualifying for a lunch subsidy. In the fall of 2013, the program served more than 51,000 children (19 percent of 3-year-olds and 29 percent of 4-year-olds) in 6% of its school districts in a 6 hours per day program, 5 days a week, and included access to wraparound services (Barnett, et al., 2015).

The Abbott program was evaluated by a series of studies that began by monitoring program quality and child readiness at kindergarten entry. This was considered essential, as the program was grown from the existing diverse delivery system, some of which was of relatively low quality when the program began. In 2000, the average ECERS-R score was 3.9. By 2008 it had risen to 5.2. Other assessments used to measure the quality of practice specific to literacy and mathematics revealed a similar pattern of low starting point and considerable progress.

Formal, evaluation of program impacts did not begin until the 2005-2006 program year. In this year, a two-pronged evaluation was launched, employing an RDD design together with a matched sample collected at kindergarten entry of children who had not attended the pre-K program, but were in the same kindergarten classrooms as the former pre-K students (the "post-test" or program group experienced pre-K the previous year, 2004-05). The evaluation found strong impacts in language, literacy, and math for one year of the pre-K program in the RDD design, while the matched groups design produced significant, but somewhat smaller estimates of the effects of one year of the Pre-K program on the same outcomes. However, the matched group design provided estimates of the effects of two years of pre-K and permitted longer term follow-up to obtain estimates of persistent effects, neither of which is possible with an RDD design. The RDD design was considered necessary to validate the matched sample design and to estimate the extent of any bias (which was clearly downward) in the matched sample design (Frede et al., 2007).

In the longer-term follow-up, the evaluation of the program found effects large enough to close at least a quarter of the achievement gap with one year and 40% with two years of pre-K through second grade (Frede & Barnett, 2011). The program also found reduced grade retention. By 4th and 5th grade the program was found to increase achievement in language arts and literacy, math, and science (Barnett, Jung, Youn & Frede, 2013). At this point, the effect of two-years of pre-K beginning at age 3 had larger persistent effects on achievement than did one year of pre-K. The gains due to one year of the program were effect sizes of 0.10 to 0.20 and the gains from two years were 0.20 to 0.40 through grade 5.

This program is characterized by smaller classes, better qualified teachers, wraparound services, and extensive supports for teacher development through an extensive continuous improvement process, all of which is much stronger than what is available in most other state pre-K programs. In fact, in the State of Preschool Yearbook for 2014, the New Jersey program ranked 2nd in spending, with 9 of 10 quality benchmarks achieved in the program (Barnett, et al., 2015). New Jersey's program is a national leader for quality standards and resources dedicated to pre-K education. Not surprisingly, the Abbott program had a much larger impact on vocabulary over one year than did the Georgia program. However, the Abbott program's estimated impacts are somewhat smaller for literacy and math than the Georgia

[3] Program resulting from a 1998 state Supreme Court ruling in Abbott v. Burke. The New Jersey Abbott Program provides voluntary pre-K for 3- and 4-year-olds in school districts where at least 40 percent of children qualified for subsidized lunch at the time of the ruling. http://www.edlawcenter.org/assets/files/pdfs/abott-v-burke/Abbott_V.pdf

program's. For the Abbott program (for most children) this is the additional effect of attending at age 4 having already had a year of quality preschool at age 3.

Oklahoma's (and Tulsa) Early Childhood Program. Oklahoma's Pre-Kindergarten was first established in 1980, and by 1998 Oklahoma had committed to provide public preschool for all 4-year-olds. The program is now available in 98 percent of school districts, serving almost 41,000 children (76 percent of its four-year olds) 2.5 to 6 hours per day, 5 days a week (Barnett, et al., 2015). Oklahoma's pre-K program is characterized by high personnel qualifications (requires a BA and a CDA for all lead teachers) matched by public school salaries; and high standards for ratio and group size. These and other standards have earned Oklahoma 8 of the 10 *State of Preschool Yearbook* for 2014 benchmarks. There have been various evaluations of Oklahoma's program, and of Tulsa in particular. The program has been found to improve cognitive development in literacy and math in Tulsa (vocabulary was not assessed), and in language as well in a statewide study. The latest evaluation yielded effect sizes at kindergarten entry that are at least as strong as those in the Abbott and Georgia studies (Gormley, Phillips & Gayer, 2008). In addition, it found Hispanic children benefited the most from the program and African-American children had strong gains, in particular when they had attended a full-day program. A very recent longitudinal study using a sample matched at kindergarten entry produced mixed results with few lasting effects detected. However, the percentage of comparison children who took the required statewide assessment at grade 3 was distinctly smaller than the percentage of former pre-K children tested, which raises questions about true comparability in the longitudinal sample.

2.2 Overview of City Programs

Boston K0-K1. With leadership of Boston's Mayor Thomas Menino in 2005, Boston Public Schools (BPS) undertook a rapid expansion to provide full day preschool for 3- (K0) and 4-year-olds (K1). BPS created the Department of Early Childhood as a first step in this initiative. The program is now available in about 85 percent of BPS locations, 6 hours a day, 5 days a week (Sachs & Weiland, 2010). Like the previous programs described, Boston's program includes high personnel qualifications (requires a BA for all staff and advancement towards an MA), and public school salaries, though with somewhat weaker requirements for group size. The first few years of the program were focused on quality improvement and professional development. Extensive data are routinely collected on classroom climate, instructional practices, and classroom organization. In addition, regular surveys are conducted with staff. This information is regularly used in coaching and supervision for program improvement. Average scores reported were 4.47 on the ECERS-R and 4.30 on the CLASS instructional support. These scores are substantially better than those in Georgia.

The most rigorous evaluation of the Boston program was carried out by Christina Weiland and Hirokazu Yoshikawa using an RDD approach. Children in the program had strong improvements in vocabulary, literacy, and math, as well as on executive function measures of working memory, inhibitory control, and cognitive flexibility (Weiland & Yoshikawa, 2013). Stronger effects were found for children from lower-income families, Latino children, and dual-language learners. The authors estimate pre-K closed the school readiness gap among poor and non-poor children in mathematics, eliminated the school readiness gap between Latino and White children in early reading and mathematics, and significantly narrowed school readiness gaps between White and Black students.

Denver Preschool Program (DPP). The DPP is a taxpayer-funded initiative (12 cent voter-approved sales tax) with the goal of increasing access to high-quality preschool. The program looks to encourage families to enroll their children in preschool through the use of credits to parents that offset tuition

costs. Tuition credits are determined by family's income, family size, and the quality rating of the center the child attends. In parallel, DPP provides funding for centers to obtain and improve their quality rating. The DPP is quite different to the previous programs described in that it does not set particular requirements for programs, but rather defines tuition credit levels that match program quality as defined by their ratings. As a consequence, within the DPP system there are significant variations in teacher qualifications, ratios, class sizes, use of curricula, and other structural factors. Outcome evaluations find significant effects on language, literacy, math, and reading through third grade (Robertson, McClelland, Palaich, Rooney and Workman, 2015). However, the initial effects are considerably smaller than those found in Georgia, New Jersey, Tulsa, and Boston, while the long-term effects appear to be based on simple comparisons of outcomes on the third grade tests between children who did and did not attend Denver pre-K, without extensive controls for possible pre-existing differences between the families of these children.

San Antonio PreK 4 SA. The San Antonio program was spearheaded when in 2011, former Mayor Julián Castro convened a task-force of Chief Executive Officers, Superintendents, and education professionals, with the mandate of identifying the most effective method for improving San Antonio's educational quality. The Taskforce recommended developing a program focused on high quality preschool for 4-year-olds. With the support of a voter-approved 1/8 cent sales tax, the program provides full-day, five days a week programming, and wraparound services, and has high teacher qualifications and high class size and ratio standards. The program started with a focus on slow growth, and processes to strengthen quality and alignment to Kindergarten, and therefore contracted an external evaluator as part of this process. So far, the program has shown growth on children's cognition, literacy, and math in a non-equivalent group comparison of children in the program to the norm (Edvance, 2014).

2.3 Program Matrix

With this previous summary in mind, we turn next to a program matrix with a description of the programs in the States of New Jersey, Oklahoma, Georgia, and in the cities of Boston, Denver, and San Antonio. As a response to the request by the City of Seattle, we identified the following set of aspects for each program, to the degree that this information was available to the research team: program goals, outcomes, and indicators; the role of the State Quality Rating and Improvement System (if applicable); teacher credential requirements and systems to support teachers in attaining these requirements; oversight and governance structure; delivery system and how public schools, community-based organizations, family child care homes, and other government subsidized programs are incorporated; child eligibility, selection process, and demographic characteristics; children served and how this has changed since program inception; tuition structure (if applicable); cost per child, funding mechanism, and teacher compensation structure; classroom structure and characteristics; curricula required or used; approach to family engagement, health support, and professional development. We summarize the program characteristics below and then provide full details in Table 2.

While State programs had large expansions in the late 1990s, city programs are newer. Most of the programs are either universal or universal within districts. City programs seemed to have grown at a slow rate, and the set of programs summarized have varying degrees of coverage. The states are in different stages of their QRIS development, and, among the city programs, only Denver seems to build on the rating system. As for validation, these do not fall under the previews of the cities, and only Oklahoma has completed a validation.

In terms of staff requirements, all of the cases summarized require a BA for teachers, except for Denver, where such requirements are quite aligned with the QRIS ratings. The programs vary in governing structures and are mostly based on mixed delivery for the three states showcased, while being in the public system in San Antonio and Boston. Child eligibility is quite universal for all these cases, with some targeting based on geographic location (either for eligibility by district or for tuition differences). Tuition is free in all except Denver and San Antonio, where there are scaled fees or subsidies. Costs per child are quite high for New Jersey and San Antonio. Funding structures do vary, with state budgets, lottery, or sales tax mechanisms (cities). Most programs are full time, with variations only in Oklahoma and Denver. Class sizes vary between 15 and 22, and wraparound services are offered only in New Jersey and San Antonio. Most programs have calendars that parallel the school year. There is wide variation in curricular requirements, professional development, and use of coaching. We also looked at the use of classroom quality measures, whether observations or checklists, and these appear to be present in most programs.

2.4 Head Start

Although information on Head Start was not requested, we include Head Start because of its significance as the nation's oldest and largest program serving disadvantaged children. The Head Start program is designed to support children's learning and development broadly so as to improve child development very broadly and thereby improve school readiness and social competence. Head Start emphasizes the provision of comprehensive services and assisting parents in their development as well as directly supporting child development. In recent years, Head Start has increased its emphasis on enhancing child outcomes, especially language and literacy; however, much of this effort came after the Impact Study children attended Head Start in 2002 (Puma et al., 2010). Head Start varied in many of its features within the study sample. Average hours of participation were 24-28 per week. Only 30 percent of the teachers had a BA degree and 30 percent an AA. Reported quality was relatively high with 70 percent rated good or better based on direct observation; however, this is at odds with the national sample of Head Start programs observed in the ECLS-B study in 2005 for which only 40 percent rated good or better based on the same observation instrument (NCES, Digest of Education Statistics, 2013).

The Head Start evaluation employed a randomized trial with 4,667 children (about half at 4 and half at age 3) attending 383 centers in 23 states. The study estimated the effects of one year's attendance at age 3 and at age 4 separately (but did not estimate the effects of 2 years). To assess this highly disadvantaged and diverse sample of children, 19 different standardized assessments were employed that took 35-45 minutes per child. These included the PPVT and TVIP and the Woodcock-Johnson and Woodcock-Munoz. Despite a huge federal budget for evaluation, child assessment instruments were not used in languages other than English and Spanish due to a lack of valid and reliable measures for these other populations in the United States. Parent ratings were used to assess children's social development and parent surveys also provided measures parents' perceptions of the child's cognitive development that can be obtained regardless of the home language. In addition to data on children, parent surveys collected extensive data on family demographics and activities. The primary classroom observation measure employed was the ECERS-R, though the Arnett and a checklist of teacher directed activities were employed, as well. Follow-up was conducted the year after attending Head Start for a year and again in Grade 3.

The Head Start Impact study provides key insights into the important of choices about evaluation methods on its own and in relation to the state and local studies. One of the most important is that randomized trials (RT) are likely to provide smaller estimates of program impacts than other

approaches, particularly the regression discontinuity design RDD). An obvious reason for this is that an RT estimates the impact relative the experiences of the control group, and the Head Start study found that some of them actually found their way into Head Start while many others attended other preschool programs (some of which were as good or better). The RDD measures the impact of a program compared to not attending a preschool program. It is also possible for an RT to create compensatory rivalry by parents denied entry who may then try to get other advantages for their children and this might help erase the initial advantage over time. Other lessons from Head Start are that it really is practical to directly measure children's learning and development in English and Spanish, while relying on parent report to provide another view for children with another home language. Finally, despite extensive analyses of subgroups no particularly interesting subgroup results emerged despite a sample size of nearly 2500 for each age group. This suggests that studies should primarily attend to questions of overall quality and impact even while looking at effects on child outcomes by ethnicity, income level, and home language.

Table 2. Program Comparison Matrix for Selected State and City Programs.

CHARACTERISTICS	STATES			CITIES		
	Georgia Bright from the Start	**New Jersey New Jersey's Abbott Preschool Program**	**Oklahoma Oklahoma's Early Childhood Program**	**Boston K0 & K1**	**Denver Denver Preschool Program**	**San Antonio Prek 4 SA**
Program Goals	Georgia's Pre-K Program is a lottery funded educational program for Georgia's four year olds to prepare children for Kindergarten. Originally started as a program to serve at-risk 4-year-olds, the program was universally opened to all 4-year-olds in September 1995. Due to budgetary constraints, enrollment has been capped in recent years to approximately 81,000 children, or 60 percent of the eligible population. [t]	In 1998, New Jersey's Supreme Court mandated that all children in the state's Abbott districts (defined as New Jersey's 30 highest poverty districts) receive a high-quality preschool education. The goal of the program is to prepare children to enter kindergarten with skills and abilities comparable to those of their wealthier suburban peers. [n]	Since 1998 Oklahoma has offered high-quality pre-K education, on a voluntary basis, to all four-year-old children. The state requires that no more than 20 children are in a class, and that child-to-staff ratios are no higher than 10 to 1. Lead teachers are required to have a B.A. and to be certified in early childhood education; they are paid according to the same salary and benefits schedule as other public school teachers. [a]	To provide all children with a strong and early start to their formal education.	To encourage families to enroll their eligible children in preschool and encourage preschool providers to improve the quality of their services. [g]	To increase the amount and quality of early childhood education across the city of San Antonio so that all children are prepared to enter kindergarten, and to be successful in school and beyond. [d]
Outcomes	Program outcomes not formally established. Goal of providing voluntary	Program outcomes have not been formally established. Goal of providing	Program outcomes have not been formally established. Goal of providing	Unknown	Program outcomes have not been formally established. Goal of providing	Program outcomes have not been formally established. Goal of providing

	access to a qualified, quality pre-K program not realized (60% served in 2013-2014). Regulations require participating programs meet 8 of 10 NIEER quality benchmarks (previously met 10; changes made in 2012 in class size and adult-child ratios no longer met standard). Child outcome goals are to prepare children for kindergarten. Child outcomes are identified in the Georgia Early Learning and Development Standards (GELDS), a research-based set of standards for what children birth - five years are capable of knowing and doing. Information on GELDS can be found at http://gelds.decal.ga.gov/Default.aspx. [a]	access to a quality pre-K program has been realized for identified districts. Regulations require participating programs to meet 9 of 10 NIEER quality standards benchmarks. Child outcome goals are to prepare children for kindergarten in five domains of development and learning. Child outcomes are identified in the Preschool Teaching and Learning Standards. (summarized from a)	universal access to a quality pre-K program has been realized for 99% of districts with 76% of children enrolled. Regulations require participating programs to meet 8 of 10 NIEER quality standards benchmarks. Child outcome goals are to prepare children for kindergarten in five domains of development and learning. Child outcomes are identified in the Early Learning Guidelines for Children available at http://www.okdhs.org/NR/rdonlyres/8D52CAF8-E29E-4C88-8131-FB9F18D46910/0/1054_EarlyLearningGuide_occs__10012010.pdf. [a]		universal access to quality pre-k program has been realized for 70% of 4-year-olds. [g]	universal access to quality pre-k program has been realized for 54% of 4-year-olds. [s]

	Georgia Bright from the Start	New Jersey New Jersey's Abbott Preschool Program	Oklahoma Oklahoma's Early Childhood Program	Boston K0 & K1	Denver Denver Preschool Program	San Antonio Prek 4 SA
Indicators	60% of children have access to pre-K. Participating programs must meet regulations which meet 8 of 10 NIEER quality standards benchmarks. Child indicators are identified in the Georgia Early Learning and Development Standards (GELDS) though results are not available. Information on GELDS can be found at http://gelds.decal.ga.gov/Default.aspx. [a]	21% of 4-year-olds and 19% of 3s have access to Abbott pre-K (2 additional state-funded programs increase the number to 30% of 4s). Participating programs must meet regulations which meet 9 of 10 NIEER quality standards benchmarks. Child indicators are identified in the Preschool Teaching and Learning Standards though results are not available. [a]	76% of 4-year-olds have access to pre-K . Participating programs must meet regulations which meet 8 of 10 NIEER quality standards benchmarks. Child indicators are identified in the Oklahoma Academic Standards (pre-K) though results are not available. [a]	Offer K1 seats in 85% of elementary schools, Early Learning Centers and K-8 schools. Also offers 'playgroups' for children 1-3 and their caregivers at some locations.	70% of 4-year-olds have access to pre-K. [g]	54% of 4-year-olds have access to pre-K. [s]
The role of the State Quality Rating and Improvement System	Quality Rated, Georgia's QRIS was launched in 2012 with programs achieving up to 3 star levels. Pre-K programs currently may participate in Quality Rated; it will likely be required in	New Jersey is currently in its second year of implementing a QRIS system with support from federal Race to the Top Early Learning Challenge funding. Grow NJ Kids, administered by	Implemented in 1998, Oklahoma's Reaching for the Stars is considered to be the first State comprehensive quality rating and improvement system. It has four levels of child care	Unknown	Uses a branch state QRIS program, Qualistar, (relation explained here: http://www.coloradoofficeofearlychildhood.com/#!quality-improvementinitiatives/chem) to evaluate preschools. It uses	N/A

(if applicable)	the near future. For additional information, see http://decal.ga.gov/ChildCareServices/QualityImprovementProgram.aspx. [u]	the Department of Human Services, has five levels. State-funded pre-K programs are not required to participate at this time.	program quality (One Star, One Star Plus, Two Star, and Three Star) and all licensed child care providers are awarded at least one star. From its inception, Reaching for the Stars has had three goals: to raise the Oklahoma Department of Human Services reimbursement rate, resulting in more slots for children whose families receive child care assistance; to improve the competency level of child care providers, in order to increase the overall quality of programs; and to provide a system whereby parents can evaluate the quality of child care program. k, [aa]		four star levels. These rankings determine the assistance preschoolers receive for attending the schools. Additional information regarding how schools are ranked at: http://www.qualistar.org/what-is-the-qualistar-rating.html.	
Evaluation/Validation of the	None available; program still in stages of early implementation. [u]	Program in early stages of implementation. NIEER will conduct an	Validation of the QRIS system in 2002 by Oklahoma State University. Examined	N/A	N/A	N/A

QRIS system		evaluation of the QRIS beginning at the end of 2015.	quality differences between various levels of the system.[ab]			
	Georgia **Bright from the Start**	**New Jersey** **New Jersey's Abbott Preschool Program**	**Oklahoma** **Oklahoma's Early Childhood Program**	**Boston** **K0 & K1**	**Denver** **Denver Preschool Program**	**San Antonio** **Prek 4 SA**
Teacher credential requirements and systems to support teachers in attaining these requirements	Lead teachers are required to have at least a bachelor's degree in early childhood education or a related field (unless previously approved). Teachers previously working with an AA will receive waivers provided they demonstrate continual progress toward a BA in their current work setting. Assistant teachers are required to have at least a Child Development Associate (CDA) credential or a paraprofessional certificate issued by the Georgia Professional Standards	Teacher required to have a BA and Certificate in P-3; Nursery school. Assistant teacher must have HSD or equivalent degree. [a]	Teachers required to have BA and EC Certification. Assistant teacher required to have HSD or equivalent. [a]	All BPS early childhood staff are required to have a bachelor's degree and to attain a master's degree within five years of entering the school system. Provide direct classroom coaching for all preschool teachers (1 coach per 10 classrooms for three years) [j]	Varies by star level: 1 star requires very little training. While at 2-3 star programs most staff has an AA degree. And at 4 star programs staff has either an AA or a BA. More detailed information at: http://www.qualistar.org/what-is-the-qualistar-rating.html	Master Teachers are degreed and accredited teachers with a minimum of three years teaching experience in early childhood classrooms. Their accreditation is EC-4 or EC – 6, or elementary with a MS in early childhood education. They have a Teacher category for TAs who have worked with us one or more years and have the appropriate degree and teaching credential. They count their years of experience as TAs as teaching experience in early childhood. They are paid at a lower salary, which is

	Commission. [a]					still comparable to public school, until they meet all the criteria as a Master Teacher. [af]
Oversight and governance structure	Bright from the Start: Georgia Department of Early Care and Learning	New Jersey Department of Education, Division of Early Childhood Education.	Oklahoma State Department of Education.	Administered by Boston Public Schools.	DPP is required to provide status reports to the Denver Office of Children's Affairs (formerly the Mayor's Office for Education and Children), a Denver city agency. A seven-member board of directors and a 25-member board of advisors oversee the program. DPP has four administrative staff: a President/Chief Executive Officer, a Senior Director of Strategy and Communications, a Program Director, and a Business Manager.	The initiative is governed by a City Council appointed 11-member board. City Council provides fiscal oversight for the initiative to include approval of the program's annual budgets Additional info at: http://www.sananto nio.gov/Pre-K4SanAntonio/About /OrganizationalChart .aspx

	Georgia Bright from the Start	New Jersey New Jersey's Abbott Preschool Program	Oklahoma Oklahoma's Early Childhood Program	Boston K0 & K1	Denver Denver Preschool Program	San Antonio Prek 4 SA
Delivery system and how public schools, community-based organizations, family child care homes, and other government subsidized programs are incorporated	Institution eligible to receive funding: Public schools, Head Start, Private CC, Faith-based centers, Other settings: Universities and technical schools, military bases, charter schools. Institutions can apply as long as they have capacity for 22 students they can apply. Application guide lines found at: http://decal.ga.gov/Prek/GuidelinesandAppendix.aspx. Programs must offer an approved curriculum. [a, m]	The New Jersey Department of Education (DOE) provides funding to eligible districts to provide the program to all resident 3- and 4-year-olds who choose to enroll. Private child care centers or Head Start programs that meet state standards may contract with districts to deliver services. Abbott districts also receive supplemental child care subsidy funds from the state Department of Human Services (DHS) to provide extended-day and extended-year services to families. [a]	Funding is allocated to public school districts which may subcontract or enter into collaborative models with Head Start, private child care, faith-based programs, and other community programs. Federally funded Head Start programs, which are targeted to poor or otherwise at-risk children and private day care centers are eligible for state funding if they establish "collaborative" relations with their local school district. http://www.sciencemag.org/content/320/5884/1723.full?ijkey=9fON.EIKV6mQA&keytype=ref&siteid=sci ; http://www.aft.org/sites/default/files/peri	Offer K1 seats in 85% of elementary schools, Early Learning Centers and K-8 schools. Also offers 'playgroups' for children 1-3 and their caregivers at some locations.	Tuition support to help families pay for pre-K, private or public. Public option is tuition based (Only for 4 year olds). [g]	Only public provision in 4 centers to date.

			odicals/BarnettFrede.pdf [a]			
Child eligibility, selection process, and demographic characteristics	Universal pre-kindergarten program for 4-year-olds; however, the number of available slots is limited resulting in programs using locally determined methods for enrollment. The program serves children from all income levels, with no fees charged to families for program participation. [a]	Universal pre-kindergarten program for 3- and 4-year-olds. The program serves children from all income levels within the 35 poorest school districts in the state--districts with at least 40 percent of children qualified for free or reduced price lunch. [a]	Universal pre-kindergarten program for 4-year-olds. The program serves children from all income levels.[a]	Lottery system. [i]	Anyone living in Denver can apply. [g]	Lottery System for those living in San Antonio in the participating districts. More detail under tuition structure. [f]

	Georgia Bright from the Start	New Jersey New Jersey's Abbott Preschool Program	Oklahoma Oklahoma's Early Childhood Program	Boston K0 & K1	Denver Denver Preschool Program	San Antonio Prek 4 SA
Children served and how this has changed since program inception	Georgia's targeted Pre-K program was established in 1993, serving 8,700 children. In 1995, the program became the nation's first state-funded voluntary universal preschool program, increasing enrollment from 15,500 4-year-olds to 44,000. Enrollment in 2014 was 81,453, down from 82,868 in 2012 yet the percent of eligible children served is currently at its highest point due to shifts in population. Enrollment has increased since the program's inception, now serving 60% of eligible children compared to 53% in 2002.[t]	As a result of a 1998 state Supreme Court ruling, the New Jersey Abbott Program provides voluntary pre-K for 3- and 4-year-olds in school districts where at least 40 percent of children qualified for subsidized lunch at the time of the ruling (31 districts). Enrollment has increased since the program's inception, now serving 30% of eligible 4s compared to 20% in 2002, and 19% of 3s, up from 11%.[c, a]	In 1980, Oklahoma began providing pre-K services for 4-year-olds on a pilot basis. In 1998, Oklahoma became the second state to offer free voluntary pre-K to all 4-year-olds.[c] Enrollment has increased since the program's inception, now serving 76% of eligible children compared to 56% in 2002. [a- c]	In his 2005 State of the City address, Boston Mayor Thomas Menino directed the Boston Public Schools (BPS) to "provide all 4-year-olds in the city with full day school within five years." The system serves 3- and 4-year-olds and is part of an initiative "Engaging families, educators and the community in a citywide effort to enhance early learning opportunities and to support successful transition into kindergarten." http://www.countdowntokindergarten.org/materials.html	580 served in 2007-2008, 5056 in 2008-2009, 5,467 in 2012-2013. Roughly 70% of 4-year-olds in Denver served. 52% of families reported incomes of less than $30,000 in 2013 (full graph available on page 18 of resource).[g]	Lottery System 700 slots in 2014, 1500 in 2015, planning on having 3,700 by 2017. http://www.sananto nio.gov/Pre-K4SanAntonio/about .aspx
Tuition structure (if	Free. [a]	Free. Abbott districts also receive supplemental child	Free. Through the state's school finance formula, public	Free.[i]	Tuition assistance: families receive aid increasing with	Four-year-olds whose family would receive free or

applicable)		care subsidy funds from the state Department of Human Services (DHS) to provide extended-day and extended-year services to families. [a,b]	school districts receive funding for the Early Childhood Four-Year-Old Program. A per-pupil rate, calculated using the age of the child and the length of the program day, is used to repay districts. Districts can support other centers by placing public school teachers in child care centers, Head Start settings, and community-based programs. Children in these sites receive the same services as children in public school locations, and are considered public school enrollees. [a,b]		family size, decreasing with family income, and increasing with preschool quality, as well as depending on if the program is part-time (At least 5 hours per week), full day (at least 25 hours per week), extended day (at least 33 hours per week); more aid is received for longer programs. [h] Preschools are ranked 1-4 stars and reevaluated every two years. Four-star system rates the quality of preschool classrooms in five areas: (1) learning environment, (2) family partnership, (3) staff training and education, (4) adult-to-child ratio, and (5) accreditation through a national accrediting agency. [g]	reduced lunch, who are unable to speak and comprehend the English language, who are homeless, who are the child of a member of an active duty the armed forces of the United States, who are the child of a member of the armed forces who was injured of killed while serving on active duty, who are or has been in the conservatorship of the Department of Family and Protective Services following an adversary hearing. For those who do not receive free access, or do not live in a participating district, there is scaling tuition: increasing with family income and decreasing with family size. [f]

	Georgia Bright from the Start	New Jersey New Jersey's Abbott Preschool Program	Oklahoma Oklahoma's Early Childhood Program	Boston K0 & K1	Denver Denver Preschool Program	San Antonio Prek 4 SA
Cost per child and how this has changed since program inception	$3,746 (state funded). This figure represents a decrease from $5,520 in 2002 when NIEER began tracking cost per child expenditures. [a]	$13,337; Increase from 2002 ($10,676) but decrease from 2010 ($12,632). [a]	$7,678 (total), $3,671 (state); Little change from 2002 ($3,601) but down significantly from 2010 ($4,885). [a]	Roughly $12,000. [i]	Families can have $10 to $419 per month of their preschool bill paid for. Due to financial constraints of the program, the maximum tuition credit awarded has been reduced from a high of $1,400 in 2007, down to $539 in 2011, $374 in 2012, and $419 in 2013.	$14,533 in 2014. $14,631 in 2015. (Transportation services, facilities lease, professional development, program assessment spending all not included) http://www.sananto nio.gov/Portals/0/Fil es/PreK4SA/FY2015 %20Annual%20Adop ted%20Budget.pdf
Funding mechanism	All state funded; Lottery revenue ($3,746 per child expenditure). [a]	All state funded; State Preschool Education Aid ($13,337 per child expenditure). [a]	State, federal, required local, and non-required local sources combine to provide total expenditures of $7,678 per child. State appropriations contribute $3,671 toward that total. [a]	Boston Public Schools	Denver voters approved a 12 cent sales tax on $100 purchases. Since 2007, over $40 million in tuition support. [g]	Voter-approved sales tax increase to fund public preschool: ⅛ cent sales tax increase to provide roughly $31 million. [d]
Teacher compensation structure	Program guidelines provide minimum salary requirements for lead teachers based on credentials, 100% of which is	Salaries for certified teachers, alternate route teachers, and teacher assistants must be set according to district	Pre-K teachers are compensated at the same level as public school teachers. Additional information may be	Teachers on same pay scale as K-12 system with same educational requirements. [i]	Varies by program	Teachers are City of San Antonio employees. Master Teacher: min. BS in elementary education EC-4 or

	funded by the state, as well as minimum salary requirements for assistant teachers meeting the credential requirements. Nonpublic schools must guarantee lead teacher SAT minimum of 90% the calculated rate. Gov. Deal's Education Reform Commission is currently examining recommendations to adjust teacher and assistant teacher salary levels. [v]	policy (as per Private Provider One Year Budget Instructions and Guidance retrieved from www.nj.gov/education/ece/archives/hs/instructions.pdf).	found at http://onlinelibrary.wiley.com/doi/10.1111/j.1541-0072.2005.00092.x/full			EC-6, 3 yrs. Early childhood teaching experience salary range: $60,057.66 - $90, 086.36. (In reality, currently at most low $70's). When only 1 year experience as a TAII with Pre-K 4 SA salary range: $49,634.26 - $74, 451.52. Teacher Assistant II. Full time co-teacher in each classroom. Min. 48 hrs. of college with coursework in ECE, 3 years experience teaching in EC, beginning salary: $12.52/hour. Teacher Assistant I: Supervises students on the bus, provides relief in classrooms and supports wherever needed. Also for extended day staff. Min. HS Diploma or GED and experience working with young children,

						beginning salary: $11.08/hour.[af]
	Georgia Bright from the Start	**New Jersey New Jersey's Abbott Preschool Program**	**Oklahoma Oklahoma's Early Childhood Program**	**Boston K0 & K1**	**Denver Denver Preschool Program**	**San Antonio Prek 4 SA**
Classroom structure and characteristics	Georgia's Pre-K Program is based on a school-year model with instruction for 6.5 hours/day during the school year. Services may be delivered in public or private programs. There is no prescribed curriculum; however, programs may select from an approved list. [a, w]	New Jersey's Former Abbott Preschool Program is based on a school-year model with instruction for 6.5 hours/day during the school year provided to all resident 3- and 4-year-old children. Programs operate in 35 districts where at least 40 percent quality for free or reduced price lunch. The mixed-model program provides services in schools, Head Start, private child care, and faith-based programs not offering religious content. There is no prescribed curriculum; however, programs may select from an approved list. [a, z]	The Oklahoma Early Childhood Four-Year-Old Program provides services 2.5 to 6 hours a day, 5 days a week during the academic year. Programs have the option of operating a part-day or a school-day program or a combination of both within each district. Districts can choose to offer a longer day but will not receive state funding for additional hours. Programs must be offered at least 175 days per year or 1080 hours per school year. [a]	Program provides services 6 hours a day. [i]	Varies by program	Program provides services 7 hours a day.[e]
Class	Maximum class size	Maximum class size	Maximum class size	Teacher and	Varies by program	Maximum class size

Size and Ratio	is 22 children; lead and assistant teacher. Staff to child ratio 1:11. Changed from 1:10 in 2012 due to budget constraints. [a]	15; lead and assistant teacher. Staff to student ratio is 2:15.[a]	is 20; lead and assistant teacher. Staff to child ratio 1:10.[a]	paraprofessional. Staff to child ratio 1:11.[i]	(Receive better ranking for better staff to child ratios).[g]	is 20; lead and assistant teacher. Staff to child ratio is 2:20. [f]
Hours of operation	6.5 hours a day, 5 days a week. [a]	6 hours a day, 5 days a week.[a]	2.5 (Part time) or 6 hours a day (Full time), 5 days a week during the academic year. Programs have the option of operating a part-day or a school-day program or a combination of both within each district. Districts can choose to offer a longer day but state funding only pays for part-time day. School day, 31,916; Part day, 8,907 [a]	6 hours a day. [i]	Increased funding for increased day length. Day length is broken into categories: part-time (At least 5 hours per week), full day (at least 25 hours per week), extended day (at least 33 hours per week).[g]	8:00 am to 3:00 pm. Afterschool care from 3:00 to 6:00 pm & Extended day care beginning at 7:15 am. 5 days per week. [ac]
Length of School Year	Parallels the school year. In 2011-2012, the number of required days was reduced by to 10% to 160 days due to fiscal constraints. In 2012-2013, days were restored to address	Parallels the school year (180 days). [a]	Parallels the school year (175 days). [a]	Parallels the school year. [ae]	Varies by program	Parallels the school year (177 days).[d]

	the resulting attrition of teachers and participating programs. [a]					
	Georgia **Bright from the Start**	**New Jersey** **New Jersey's Abbott Preschool Program**	**Oklahoma** **Oklahoma's Early Childhood Program**	**Boston** **K0 & K1**	**Denver** **Denver Preschool Program**	**San Antonio** **Prek 4 SA**
Curricula required or used	Guidelines for classroom instruction are provided through Georgia's Pre-K Program Content Standards ii, which are aligned with Georgia's Early Learning Standards and Georgia's Kindergarten Performance Standards. The program standards also require Georgia's Pre-K sites to use an approved curriculum; provide written lesson plans which include educational experiences in language/literacy, math, science, social studies, creative (music, art, and drama), social and	Curricula in pre-K programs must align with the Preschool Teaching and Learning Standards available at http://www.nj.gov/education/ece/guide/standards.pdf. Further, the Office of Early Childhood Education (OECE) supports preschool curricula that meet the following criteria: - The curriculum is aligned with the Preschool Teaching and Learning: Expectations: Standards of Quality; - Methods for inclusion of students with disabilities are provided; - The content and teaching strategies	Rules and regulations for early childhood education programs in the 2012-2013 Standards for Accreditation of Oklahoma Schools state, "The curriculum shall be appropriate for the age and developmental level of the students. A process to provide continuity between the early childhood program and the kindergarten program shall be established." No specific curriculum model is identified for all programs to adopt; however, curriculum should address the PASS Standards for Pre-	Opening the World of Learning (OWL), by J.A. Schickedanz and D. Dickinson and the mathematics curriculum (Real Math Building Blocks: PreK, by D.H. Clements and J. Sarama) for all preschool programs. [i]	Varies by program	Engaging lessons with Frog Street Press and Teaching Strategies Curricula http://www.sanantonio.gov/Pre-K4SanAntonio/Facts.aspx

	emotional, and physical development; implement individual child assessments using the Georgia's Pre-K Child Assessment—Work Sampling Online, which is based on the Work Sampling System [t, w]	are clear, and research-based; - The curriculum content is taught with focus and integration; - The curriculum relies on child initiation and engagement; - All curriculum components are developmentally appropriate; and - Show evidence of benefits. The recommended curricula are: The Creative Curriculum®, Curiosity Corner®, HighScope Preschool Curriculum, Tools of the Mind. [z]	Kindergarten (available at http://ok.gov/sde/sites/ok.gov.sde/files/documents/files/PASS_PreK_OSDE.pdf).			

	Georgia Bright from the Start	New Jersey New Jersey's Abbott Preschool Program	Oklahoma Oklahoma's Early Childhood Program	Boston K0 & K1	Denver Denver Preschool Program	San Antonio Prek 4 SA
Approach to family engagement	Family engagement is an integral part of the Pre-K program. DECAL developed a Pre-K Family Handbook to inform and encourage family engagement, and it is an important component of the QRIS system. DLL support is also available to programs serving families requiring such services. [x, y]	Support services are provided to families (see below). Families for whom English is not a primary language are provided the following additional supports: a home language survey is sent home at the beginning of the school year; all Information must be presented to parents in their primary language; and translators or bilingual staff are available if children do not speak English. [a]	Support services are provided to families (see below). Families for whom English is not a primary language are provided the following additional supports: a home language survey is sent home at the beginning of the school year; information must be presented to parents in their primary language; and translators or bilingual staff are available if children do not speak English. [a]	Unknown	Varies by program	Sixty-four family events were held in the first half of the year with an average attendance of almost 26 (25.6) individuals. Eighty-eight events were held in the second half of the year with an average attendance of nearly 22 (21.9) individuals.[d]
Support Services	The following support services are required to be provided by programs: parent involvement activities; nutrition information to families; referral for	The following support services are required to be provided by programs: parent support and training; parent involvement activities; nutrition information to	The following support services are required to be provided by programs: parent involvement activities; nutrition information to families; referral for	Unknown	Varies by program	The following support services are present: breakfast, lunch and 2 snacks, workshops to families, transportation, family specialists supporting families,

	social services; transition to kindergarten activities; parent conferences and/or home visits; with support services determined locally. [a] Offers meals, rest time, and both indoor and outdoor play time; and provide support services or referrals to families as needed. [t, w]	families; referral for social services; transition to kindergarten activities; parent conferences and/or home visits. [a]	social services; transition to kindergarten activities; and parent conferences and/or home visits. [a]			information sessions on child development, health and education. [ac]
Professional Development	Teacher in-service requirements 15 clock hours per academic year. [a]	Teacher in-service requirements 100 clock hours per 5 years. [a]	None. Since the 2010-2011 school year, professional development has not been required by the state. As per Oklahoma HB 2928, "A licensed or certified teacher shall not be required to complete any points of the total number of professional development points required." It is a local school district's decision how many	Provides direct classroom coaching for all preschool teachers (1 coach per 10 classrooms for three years). Ensures consistent and targeted professional development. Offer a professional early childhood fellowship for elementary school principals. [j]	Varies by program	Professional development is a major focus of Pre-K 4 SA. Fourteen coaches with backgrounds ranging from child development to bilingual studies and special education are available to collaborate with partner school districts to impact early childhood education in San Antonio. [ad]

			hours of professional development are required. [a]			

	Georgia Bright from the Start	New Jersey New Jersey's Abbott Preschool Program	Oklahoma Oklahoma's Early Childhood Program	Boston K0 & K1	Denver Denver Preschool Program	San Antonio Prek 4 SA
Use of Classroom Observation Measures	Structured observations of classroom quality (CLASS, annually); Documentation of children's learning and/or child outcomes (WSS rating every six weeks, report to parents twice a year); Review of program facilities and safety procedures (Child Care License report); Results of program self-assessments (Grant Requirement Checklist, IQ Guides for Classroom Environment, Daily Schedule, Assessment); Review of program records (Grant Requirement Checklist).[a]	Structured observations of classroom quality (locally determined, most districts use ECERS); Documentation of children's learning and/or child outcomes (High/Scope's COR, Teaching Strategies GOLD, ELAS, locally selected tools); Documentation of program-level outcomes (Informal data tool collected annually by the Department of Education); Review of program facilities and safety procedures (During annual SAVS); Results of program self-assessments (During annual SAVS); Review of program records (During annual SAVS and also in annual	Documentation of children's learning and/or child outcomes (locally determined outcomes aligned with student standards); Documentation of program-level outcomes (locally determined outcomes aligned with student standards); Review of program facilities and safety procedures (Regional Accreditation officers review); Review of program records (Regional Accreditation officers review).[a]	ECERS–R (Early Childhood Environmental Rating Scale–Revised), which focuses on the global and the structural quality of the classroom; CLASS (Classroom Assessment Scoring System), which focuses on process characteristics of the classroom, including the richness of interactions between children and teachers; ELLCO (Early Language and Literacy Classroom Observation Tool), which measures the quality of the literacy environment.	Unknown	Structured observations of classroom quality using CLASS, SNAPSHOT and TSEEQ [d]

		program plan review).[a]				

REFERENCES

a Barnett, W.S., Carolan, M.E., Squires, J.H., Clarke Brown, K., & Horowitz, M. (2015). *The state of preschool 2014: State preschool yearbook*. New Brunswick, NJ: National Institute for Early Education Research.

b Peisner-Feinberg, E. S., Schaaf, J. M., LaForett, D. R., Hildebrandt, L. M., & Sideris, J. (2014). Effects of Georgia's Pre-K Program on Children's School Readiness Skills: Findings from the 2012–2013 Evaluation Study. *Frank Porter Graham Child Development Institute, University of North Carolina at Chapel Hill*.

c Wong, V. C., Cook, T. D., Barnett, W. S., & Jung, K. (2008). An effectiveness-based evaluation of five state pre-kindergarten programs. *Journal of Policy Analysis and Management*, *27*(1), 122-154.

d Pre-K 4 SA Evaluation Report.(2013). San Antonio, Texas: Edvance Research.

e Learn, Grow, Succeed. (2014). Pre-K 4 SA An Investment In Education City of San Antonio, Retrieved May 29, 2015 from http://www.sanantonio.gov/Portals/0/Files/prek4sa/factsheetApril.pdf.

f Pre-K 4 SA – Summary Fact Sheet. (2012, August 9). Retrieved May 29, 2015, from https://www.sanantonio.gov/portals/0/Files/Summary Fact Sheet.pdf.

g An Evaluation Of The Denver Preschool Program 2012-2013. (2013). Denver, Colorado: Augenblick, Palaich and Associates,Inc.

h Tuition Calculator. (n.d.). Retrieved May 29, 2015, from http://www.dpp.org/calculator.

i Yoshikawa, H. (2014, October). Investing in Our Future: The Evidence Base on Quality Preschool Education. Symposium on Children. Lecture conducted from Crane Center for Early Childhood Research and Policy, Columbus.

j Sachs, J. A. & C. Weiland. (2010). Boston's Rapid Expansion of Public School-Based Preschool. *Young Children*, 75.

k Quality Rating & Improvement System Resource Guide. (n.d.). Retrieved May 29, 2015, from https://occqrisguide.icfwebservices.com/index.cfm?do=qrisstateinfo&stateID=86

l Gormley, W., & Phillips, D. (2005). The Effects of Universal Pre-K in Oklahoma: Research Highlights and Policy Implications. *Policy Studies Journal,* 65-82.

m Becoming a Georgia's Pre-K Program Provider. (2015). Retrieved June 4, 2015, from http://decal.ga.gov/BftS/StartingAGeorgiaPrekProgram.aspx?source=green#Who.

n Barnett, S., Robin, K., Hustedt, J., & Schulman, K. (2003). The state of preschool: 2003 state preschool yearbook. New Brunswick, NJ: NIEER.

o Hill, C. J., Gormley Jr, W. T., Adelstein, S., & Willemin, C. (2012). The Effects of Oklahoma's Pre-Kindergarten Program on 3rd Grade Test Scores. Center for Research on Children in the United States, Georgetown University, Washington, DC, 19.

p View State Profiles. (2014). Retrieved May 29, 2015, from http://qriscompendium.org/view-state-profiles.

q Academics & Strategic Direction. (2015). Retrieved May 29, 2015, from http://www.Bostonpublicschools.org/Page/279.

r Boston Public Schools: Overview. Retrieved May 29, 2015.

bpsearlychildhood.weebly.com/uploads/1/0/1/3/10131776/boston_k1ds_overview.pdf.
s Early Childhood Education Initiative. (2014). Pre-K 4 SA An Investment In Education. City of San Antonio .
t Bright from the Start Georgia Department of Early Care and Learning. (2015). Retrieved May 29, 2015, from http://decal.ga.gov/Prek/20YearAnniversary.aspx.
u DECAL - Quality Rating and Improvement System. (2015). Retrieved May 29, 2015, from http://decal.ga.gov/ChildCareServices/QualityImprovementProgram.aspx; also interview with Susan Adams (DECAL).
v FY2015 Georgia's Pre-K Rates. (2015, February 25). Retrieved May 29, 2015, from http://decal.ga.gov/documents/attachments/RatesGovernorsrecommendation2015.pdf ; also, interview with Susan Adams (DECAL).
w Georgia Pre-K Program Approved Curriculum List 2014-2015. (2015). Retrieved May 29, 2015, from http://www.decal.ga.gov/documents/attachments/GAPreKApprovedCurriculumList.pdf.
x Georgia's Pre-K Program: Family Handbook. (2013). Retrieved May 29, 2015, from http://www.decal.ga.gov/documents/attachments/PreKFamilyHandbook.pdf.
y A Roadmap for Engaging Families with Young Children. (n.d.). Retrieved May 29, 2015, from http://www.decal.ga.gov/documents/attachments/EarlyChildhoodGa-web.pdf.
z Early Childhood Education. (2010). Retrieved May 29, 2015, from http://www.state.nj.us/education/ece/curriculum/curriculum.htm.
aa DHS - Reaching for the Stars Service Information. (2014, November 13). Retrieved May 29, 2015, from http://www.okdhs.org/programsandservices/cc/stars/ also, correspondence from Kayla Hindman, Oklahoma Dept. of Education.
ab Norris, D., Dunn, L., & Eckert, L. (2003). “Reaching for the Stars” Center Validation Study Final Report. Retrieved May 29, 2015, from http://www.okdhs.org/NR/rdonlyres/4E9FC113-3988-4B31-88F6-3C5E9C993B52/0/ReachingForTheStarsCenterValidationStudyFinalReport_dcc_05212007.pdf.
ac Pre-K San Antonio: Facts. (2014) Retrieved May 29, 2015. http://www.sanantonio.gov/Pre-K4SanAntonio/Facts.aspx.
ad Pre-K San Antonio: For Teachers. (2014). Retrieved May 29, 2015, from http://www.sanantonio.gov/Pre-K4SanAntonio/Teachers.aspx.
ae Preschool (K0-K1). (2015). Retrieved May 29, 2015, from http://bpsearlychildhood.weebly.com/preschool-k0-k1.html.
af Correspondence with Kathleen Bruck (PRE-K 4 SA).

3. Evaluating programs and Instruments.

In addition to understanding the programs themselves, the City of Seattle requested detailed information regarding how the programs described above, and similar programs, have been evaluated; aspects related to communication of results and causality as linked to such evaluations; and aspects related to measures of quality of care and for children's development over time. This section addresses these four requests.

3.1 Impact Evaluation Strategies for selected States and Cities.

In relation to program evaluation, the City of Seattle requested information on research questions; data collection protocol; research methods; and significant findings. Key information on evaluations of the programs discussed above has already been provided in the text. Table 3 below describes these and other important State and City pre-K impact evaluations in the United States carried out since 2000. The table includes information on the year of the study, the research questions, the study design, follow-ups, instrumentation used in the evaluation, sample sizes, and main findings for each study. This table can be used to compare specific aspects of evaluation strategies across all of the studies.

The most robust research method, which provides the most accurate estimates, is a randomized control trial. Randomized control trials are the gold standard for impact evaluation. A randomized control trial is the most trusted 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 who attend preschool and children who do not attend preschool. This approach works best when acceptance to the program depends on a lottery system and when it is possible to retain nearly all of the participants (especially the controls who do not attend the program) in the study at follow-up. This design provides the greatest confidence that the treatment and comparison groups do not differ on either measured or unmeasured characteristics at baseline. It also provides the greatest statistical power for any given sample size. For example, even with just over 200 participants in the pilot of universal pre-K in Providence, Rhode Island, a randomized trial was able to detect moderate sized effects overall and for subgroups by income level. The randomized trial was used somewhat less successfully in Tennessee (at least up to third grade when the statewide test is administered) because of a high refusal rate in the control group. Possibly, the very large sample size in Tennessee actually reduced the researchers' ability to follow-up effectively with the entire sample.

Relevant evidence on State pre-K programs also comes from studies using what is known as a regression discontinuity design. 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 approach addresses the problem of selection bias, by comparing children just entering the preschool program (control group), and children who have just finished the preschool program and are currently beginning the 4-year-old preschool year or kindergarten (the treatment group). The RDD methodology uses a stringent, specified age cut-off for preschool eligibility to define the treatment and

control groups among the children in the study, as this cut-off is external to families. In essence, children who attend the program one year are compared to slightly younger children who must wait to attend the following year. Studies using this methodology have found gains in language, literacy, and mathematics in state pre-K programs sometimes even analogous to those found in the Chicago Longitudinal Study (Barnett, 2011; Barnett, et al., 2013; Gormley et al., 2008; Hill, Gormley, & Adelstein, 2012; Peisner-Feinberg, Schaaf, LaForett, Hildebrandt, & Sideris, 2014; Peisner-Feinberg & Schaaf, 2011; Wong et al., 2008). Given that when state pre-K programs have been studied simultaneously with both regression discontinuity and other methods (including in one instance a randomized trial) and these methods have been consistent (Frede et al., 2009; Lipsey et al., 2011), outcomes found are quite robust. RDD designs require much larger samples than RCT design methods to be able to detect similar effect sizes. They also have been criticized by those suspicious of the large effect sizes found by many RDD studies. However, it should be noted that some RDD studies have found small or even zero effects for some outcome measures, suggesting that they are not significantly biased upward.

A third type of design that has also been used in the evaluations summarized below is propensity score matching. This type of design is, in essence, a comparison of groups that are not equivalent, but uses statistical matching procedures to balance the comparison groups to each other as much as possible. In this design, a comparison group is formed by matching children in the treatment group to children outside of the program based on socio-economic characteristics.
Propensity score matching is a statistical technique used to create a group of children who are similar (probabilistically) to the treatment children based on variables of interest (using potentially stable characteristics, such as gender, race, age, and location, among others). When done well, matching can produce reasonable estimates (Shadish, Cook, and Campbell, 2002). Propensity score matching designs also require large samples, and allows for longitudinal comparisons.

Other non-equivalent group comparisons use pre- and post-tests or only post-tests. These designs are weaker, but sometimes are the only approaches feasible. While inspiring less confidence, these approaches are easily implemented to at look students over time (longitudinally). They are stronger if they can have very local matches (students from the same neighborhoods or even the same kindergarten classes, for example) and if they have pre-tests or detailed measures of children's family backgrounds. If comparison children are selected and assessed before the beginning of the preschool program, these designs are stronger. This requires either a waiting list for entry into the program or access to children attending some other type of preschool program, as it tends to be very difficult to find comparison children who are at home. Most longitudinal evaluations of program effects that compare to other children or to a standardized sample in a measure are of this type of design.

Table 3. Descriptions for State and Local Pre-K Evaluations.								
State or City	Year Study Began	Research Questions	Research Design	Data Collection Protocol			Summary of Findings	Reference
				Age of Follow-up	Measures	Sample Size (children)		
STATES								
Arkansas	2006	What is the effect of the state-funded preschool education program on children's learning at the beginning of kindergarten? (Difference in language, literacy, and mathematics test scores).	RDD	Pre-K	PPVT (language) & TVIP (language in Spanish) /Pre-CTOPPP (literacy, in English and Spanish) and WJ (math) and WM (math in Spanish).	901	Significant effects on language, literacy, & math.	Barnett, et al.(2013)
	2005	(1) What is the impact of ABC regardless of whether the comparison group attended another program (including Head Start and private preschool)? (2) What is the impact of ABC compared to not attending any center-based preschool at age 4?	RDD and "within-cohort" longitudinal.	First through 4th Grade.	PPVT (language) & TVIP (language in Spanish) /Pre-CTOPPP (literacy, in English and Spanish) and WJ (math) and WM (math in Spanish).	1,600	Effects on language, literacy & math after pre-k; on literacy after 3rd grade.	Jung, Barnett, Hustedt, & Francis (2013)
California	2006	What is the effect of the state-funded preschool education program on children's learning at the beginning of kindergarten? (Difference in language, literacy, and mathematics test scores).	RDD	Pre-K	PPVT (language) & TVIP (language in Spanish) /Pre-CTOPPP (literacy, in English and Spanish) and WJ (math) and WM (math in Spanish).	1,630	Significant effects on language, literacy & math.	Barnett, et al.(2013)
Florida	2002	Determine the effects of student	Non-	Pre-	Family and Child	3,450	Significant effect on	King, Cappellini, &

	and family characteristics on school readiness pretest scores. Determine the extent to which school readiness post-test scores are dependent on program characteristics. To determine whether beginning kindergarten students who participated in school readiness programs have higher mean scores on uniform screening instruments, show higher levels of performance, lower proportions of retention in grade.	equivalent group comparison: Control made up of program eligible non-attendees who attended no other form of pre-K.	K-Grade 3	Experience Survey (FACES): includes PPVT, Letter-Word Identification, Applied Problems, Dictation Tasks from Woodcock-Johnson Psycho-Educational Battery (WJ), Story and Print Concepts, McCarthy Scales of Children's Abilities subtest, Phonemic Analysis subtest of the Test of Language Development, Color Naming and Counting task (developed especially for FACES), Social Awareness items from the Comprehensive Assessment Program (CAP).	estimate	overall development in Pre-K, K. Significant effects on retention in K, unclear results afterwards. Significant effects on reading and math test scores in K, no lasting effects afterwards. Mixed results on attendance in K, unclear results afterwards.	Rohanic (1995)

State or City	Year Study Began	Research Questions	Research Design	Age of Follow-up	Measures	Sample Size (children)	Summary of Findings	Reference
Georgia	2012-13	1. Does participation in Georgia's Pre-K Program improve children's school readiness skills (language, literacy, math, general knowledge, behavior) compared to children who have not attended the program? 2. Are the effects of Georgia's Pre-K Program on school readiness skills similar for different groups of children on the basis of family income, gender, or children's level of English language proficiency?	RDD	Pre-k	WJ (picture vocabulary, Letter-Word Identification, sound awareness, word attack, and applied problems) /Social Awareness Task (general knowledge)/ SSIS (social skills and behavior problems)/ Counting Task/ Letter Knowledge.	1,181	Significant effects on literacy, math, and general knowledge. No significant effects on vocabulary or behavior skills.	Peisner-Feinberg, Schaaf, LaForett, Hildebrandt, & Sideris (2014)
Michigan	2004	To estimate the effects of state-level pre-K programs.	RDD	Pre-K	PPVT (language) & TVIP (language in Spanish) /Pre-CTOPPP (literacy, in English and Spanish) and WJ (math) and WM (math in Spanish).	871	Significant effects on early math and print awareness in pre-K. No effects on vocabulary.	Wong, et al. (2008)
	2004	The research question of interest is whether attendance in the state-funded preschool program at age 4 has an impact on children's academic skills at kindergarten entry.	RDD	Pre-K	PPVT (Language), Woodcock-Johnson Tests of Achievement (Math), the Blending subtest of the Preschool	865	Significant effects in language and math.	Lamy, Barnett, & Jung (2005)

					Comprehensive Test of Phonological & Print Processing (Pre-CTOPPP) (Language).			
	1996	The purpose of the MSRP evaluation is to assess how well the program contributes to, and helps parents contribute to, children's development and readiness for school participation.	PSM - Longitudinal	Pre-K, Post-test only	MSRP Program Quality Assessment, High/Scope Child Observation Record, Child Development Rating, (Initiative, Social Relations, Creative Representation Music and Movement Lang. & Literature, Logic & Mathematics).	21,077	Significant effects on Child Development Rating and Child Observation Record.	Florian, Schweinhart, & Epstein (1997)
North Carolina	2003	What were the key characteristics of the local More at Four programs and to what extent have they changed over time? What was the quality of the More at Four pre-k and kindergarten programs attended by children? What were the longitudinal outcomes from pre-k through kindergarten for children who attended the More at Four Program? What factors were associated with better outcomes for children?	RDD	Pre-K	Early Childhood Environment Rating Scale-Revised (ECERS-R), the Early Language and Literacy Classroom Observation (ELLCO), the Caregiver Interaction Scale (CIS), and the Assessment of Practices in Early Elementary Classrooms (APEEC; used only in the kindergarten year).	992	Significant effects in language and literacy skills, pre-math skills, and general knowledge.	Peisner-Feinberg & Shaaf (2008)

State or City	Year Study Began	Research Questions	Research Design	Age of Follow-up	Measures	Sample Size (children)	Summary of Findings	Reference
North Carolina	2002	What were the characteristics of the local programs? Who was served by the More at Four Program? What was the quality of the services provided? How satisfied were families with the More at Four Program? What were the outcomes of children attending the More at Four Program? What factors were associated with better outcomes for children?	Pre Post	Pre-K	PPVT-III, WJ-III, Naming Letters, Story and Print Concepts, (Language and literacy). WJ-III Applied Problems and counting task (Math). Social awareness and color naming (general knowledge). Social Skill Rating System (SSRS) social skills and SSRS problem behaviors (Classroom behavior).	6,125	Significant effects in language and literacy skills, math skills (Counting task), general knowledge, classroom behavior (SSRS Problem Behavior). No significant effect in math (WJ-III Applied Problems) and Classroom behavior (SSRS Problem Behaviors).	Peisner-Feinberg & Maris (2005)
	2009	Does participation in the More at Four Pre-k Program improve children's language/literacy and math school readiness skills? Are the effects of More at Four on school readiness skills similar for different groups of children on the basis of poverty status, English language proficiency, or cumulative risk?	RDD	Pre-K	PPVT (Literacy), Test of Preschool Early Literacy (Literacy), Woodcock Johnson-III (WJ-III) (Math), (National Center for Early Development and Learning) NCEDL Counting Number Task (Counting).	1,010	Significant effects in math and counting. No significant effects on literacy.	Peisner-Feinberg & Schaaf (2011)
New Jersey	2004	To estimate the effects of state-level pre-K programs.	RDD	Pre-K	PPVT (language) & TVIP (language in Spanish) /Pre-CTOPPP	2,075	Significant effects on vocabulary, early math, and print	Wong, et al. (2008)

				(literacy, in English and Spanish) and WJ (math) and WM (math in Spanish).		awareness in pre-K.	
2005	Determine what the short-term, yearly, and long-term effects of the Abbott Preschool program are on children's academic skills from early in kindergarten through fourth grade and whether these children are less likely to have been retained in grade or placed in Special Education than children who did not attend.	RDD with a Longitudinal Cross Sectional Design.	Pre-K, and First to 4th Grade.	PPVT (language) & TVIP (language in Spanish) /Pre-CTOPPP (literacy, in English and Spanish) and WJ (math) and WM (math in Spanish).	RDD = 1,544 - Longitudinal = 1,012	Significant effects on vocabulary, early math, and print awareness in pre-K. Stronger effects for 2 years of pre-K. Significant effects on grade retention. The 4th and 5th grade follow-up found increased achievement in Language and Literacy, Math, and Science (larger for two years of pre-K), and decreased grade retention and special education placement rates (without differences by dosage).	Frede, Jung, Barnett, Lamy & Figueras (2007); Frede, Jung, Barnett, & Figueras (2009) and Barnett, Jung, Youn & Frede (2013)

State or City	Year Study Began	Research Questions	Research Design	Age of Follow-up	Measures	Sample Size (children)	Summary of Findings	Reference
Oklahoma	2004	What is the effect of the state-funded preschool education program on children's learning at the beginning of kindergarten? (Difference in language, literacy, and mathematics test scores).	RDD	Pre-K	PPVT (language) /Pre-CTOPPP (literacy) WJ (math).	836	Significant effects on language, literacy & math.	Barnett, et al. (2013) - update from the Wong, et al (2008) analyses below.
	2004	To estimate the effects of state-level pre-K programs.	RDD	Pre-K	PPVT (language) /Pre-CTOPPP (literacy) WJ (math).	838	Significant effects on vocabulary and none on early math, and print awareness in pre-K.	Wong, et al. (2008)
	2004	Whether attendance in the state-funded preschool program at age 4 has an impact on children's academic skills at kindergarten entry.	RDD	Pre-K	PPVT (language) /Pre-CTOPPP (literacy) WJ (math).	838	Significant effects in children's vocabulary (equivalent to four months), early math, and in understanding of print concepts. No significant effects on a measure of children's skills in phonological awareness.	Lamy, Barnett, & Jung (2005)
South Carolina	2004	What is the effect of the state-funded preschool education	RDD	Pre-K	PPVT (language) & TVIP (language in	777	Significant effect on print awareness. No	Barnett, et al.(2013)

		program on children's learning at the beginning of kindergarten? (Difference in language, literacy, and mathematics test scores).			Spanish) /Pre-CTOPPP (literacy, in English and Spanish) and WJ (math) and WM (math in Spanish).		significant effect on vocabulary.	
	2004	To estimate the effects of state-level pre-K programs.	RDD	Pre-K	PPVT (language) & TVIP (language in Spanish) /Pre-CTOPPP (literacy, in English and Spanish) and WJ (math) and WM (math in Spanish).	777	Significant effects on print awareness. No effects on vocabulary.	Wong, et al. (2008)
Tennessee	2009	1. Does participation in TN-VPK improve the school readiness of the economically disadvantaged children eligible for the program? 2. What are the characteristics of the children who benefit the most from TN-VPK?	RCT and RDD (the latter not reported yet).	Pre-K	WJ (Picture vocabulary, Letter-Word Identification, Spelling, Oral Comprehension, Applied problems and Quantitative Concepts)/ Academic Classroom and Behavior Record / Cooper-Farran Behavioral Rating Scales (Social skills and behavior).	RCT = 1,077	Statistically significant on the composite measure as well as on each of the WJ literacy, language, and math measures with effect sizes between .12 and .46. The largest effects were on Letter-Word Identification and Quantitative Concepts.	Lipsey, Hofer, Dong, Faran, & Bilbrey (2013)

State or City	Year Study Began	Research Questions	Research Design	Age of Follow-up	Measures	Sample Size (children)	Summary of Findings	Reference
Texas	1991	Evaluating if prekindergarten in Texas has a positive effect on academic achievement as measured by standardized test scores and school progress as measured by retention in grade and placement in special education. Evaluating if structural quality dimensions also have a statistically significant positive effect on measures of academic success that also vary by sub-population.	Non-equivalent group comparison.	Grade 3-Grade 8	TASS reading scores, TAAS math scores.	682,749 (pooled cohorts)	Significant effect of pre-K in reading and math, and school progress, measured by grade retention and placement in special education.	Kuhne (2008)
West Virginia	2004	What is the effect of the state-funded preschool education program on children's learning at the beginning of kindergarten? (Difference in language, literacy, and mathematics test scores).	RDD	Pre-K	PPVT (language) & TVIP (language in Spanish) /Pre-CTOPPP (literacy, in English and Spanish) and WJ (math) and WM (math in Spanish).	720	Significant effect on print awareness. No significant effect on vocabulary or math.	Barnett, et al. (2013)
	2004	To estimate the effects of state-level pre-K programs.	RDD	Pre-K	PPVT (language) & TVIP (language in Spanish) /Pre-CTOPPP (literacy, in English and Spanish) and WJ (math) and WM (math in Spanish).	720	Significant effect on print awareness.	Wong, et al. (2008)

CITIES								
Boston	2008	What is the impact of the prekindergarten program on children's early mathematics, language, literacy, EF, and emotional development? Do some child subgroups (as defined by family income, race or ethnicity, or child gender) benefit statistically significantly more from the prekindergarten program than others?	RDD	Pre-K	PPVT (language) /WJ (LW & AP; literacy & math)/ERQ & TOQ & FDS & BDS & Pencil Tapping & DCCS (emotional development & EF skills).	2,018	Significant effects on language, literacy, math, emotional development, and EF skills.	Weiland & Yoshikawa (2013)
Chicago	1986 entered kindergarten	1. Does participation in the Child–Parent Center preschool program affect measures of cognitive and non-cognitive skills above and beyond child and family background factors? 2. To what extent, both separately and together, do measures of cognitive and non-cognitive skills account for the estimated direct effects of Child–Parent Center preschool participation on high school completion and incarceration in young adulthood?	PSM	Kindergarten& ages 12, 14, 24	ITBS reading score, ITBS math score, Classroom adjustment, troublemaking behavior, social skills, any delinquency.	1,539	Significant effects on language, literacy, math.	Reynolds, Temple, & Ou (2010)

State or City	Year Study Began	Research Questions	Research Design	Age of Follow-up	Measures	Sample Size (children)	Summary of Findings	Reference
Chicago, continued	1986	Is any participation in the program associated with school performance? Is duration of participation associated with children's scholastic performance? Does participation in extended intervention lead to better school performance than less extensive intervention? Which factors mediate the effects of participation?	PSM	K – 8th Grade age	Iowa Tests of Basic Skills (ITBS) standard scores.	1,150	Literacy, math, life skills.	Reynolds (1997)
Denver	2008	To determine if there is a persistence of preschool effects on third-grade academic performance.	PSM	Pre-K	TCAP Reading, TCAP Math, TCAP Writing, Developmental Reading Assessment (DRA).	Matched DPP = 2,431, Sample Non-DPP = 2,496	Significant effects on language, literacy, math, and reading.	Robertson, McClelland, Palaich, Rooney and Workman (2015)
Rhode Island	2009	What are the effects of enrollment in these programs on children's early learning outcomes? What is the effect of classroom quality on children's outcomes? Do effects of participation differ for low- and middle-income children? Do effects of classroom quality differ for low- and middle-	Randomized Control Trial.	Pre-k	PPVT (language) /WJ AP (math) /TOPEL, HTKS.	242	Significant effects in math and language. Stronger effects for low income children.	Francis, J. (2011a,b)

		income children? How does pre-K classroom quality vary when different inputs are present?						
San Antonio	2013	5a.) Is the Pre-K 4 SA program associated with a change in Pre-K 4 SA children's GOLD outcomes at the end of Pre-K 4 SA? How do Pre-K 4 SA children compare to a nationally representative normed sample of children? 5b.) Do differences in findings exist based on child characteristics, the area of readiness for kindergarten, or location children attended (North or South center)?[1]	Non-equivalent group comparison (to the normed sampled).	Pre-k	Teaching Strategies GOLD (cognitive, literacy, mathematics, oral language, physical, socio-emotional).	75.1% of children in the program; n = 555	Higher than the norm on cognition, literacy and math. At the norm on oral language, physical, and social-emotional development.	Edvance (2014)
Tulsa, OK	2006	The objective of this work is to determine how much Hispanics benefit from a high-quality pre-K program and which Hispanic students benefit the most.	RDD	Pre-K	WJ Letter-Word Identification Test, Spelling Test,(language and literacy) / Applied Problems Test (math).	Pre-K alumni = 194, Pre-K entrants = 245	Significant effects in reading skills, writing skills, and math skills.	Gormley, et al. (2008)

State or City	Year Study Began	Research Questions	Research Design	Age of Follow-up	Measures	Sample Size (children)	Summary of Findings	Reference
	Early Cohort: 2000 (2000 and 2005 pre-K cohorts)	Investigates the persistence of short-term effects of a high-quality school-based pre-kindergarten program.	PSM	Pre-K & 3rd grade	Oklahoma Performance Index (OPI).	Treatment group = 1,038, Control group = 961	Significant effects on cognitive and language after pre-K. No significant effect on reading OPI and math OPI. No significant effect on cognitive, language, reading OPI, math OP in third grade.	Hill, Gormley, & Adelstein, (2012)
	Late Cohort: 2005 (2000 and 2005 pre-K cohorts)	Investigates the persistence of short-term effects of a high-quality school-based pre-kindergarten program.	PSM	Pre-K & 3rd grade	Oklahoma Performance Index (OPI).	Treatment group = 2,024, Control group = 1,087	Significant effects on Letter Word ID, Spelling, Applied Problems after pre-K. No significant effect on reading OPI or math OPI after pre-K. Significant effects on math OPI in third grade. No significant effect on Letter Word ID, Spelling, Applied Problems reading OPI in third grade.	Hill, Gormley, & Adelstein, (2012)

[1]Evaluation also included other components related to program participation, attendance, engagement and program quality.

3.2. Communication to the public, causality, and course corrections in the evaluation process

The City of Seattle also requested information on methods used to communicate results of State and City pre-K evaluations in the United States with the public, oversight bodies, families, teachers, providers, and administrators. Further, the City of Seattle is interested in understanding when State and City pre-K evaluations began producing positive results that were causally linked to the program, as well as any course corrections to the pre-K programs that were necessary in the early years of the program. Table 4 highlights the limited information available in these areas for three State (Georgia, New Jersey, and Oklahoma) and three City (Boston, Denver, and San Antonio) pre-K programs.

Virtually no information is included in published reports that explicitly outlines approaches used by States or Cities to disseminate evaluation strategies or results with any groups of stakeholders. The City of San Antonio did hire both a public relations/marketing contractor and then a full-time Director of Communications to coordinate communications of program information to key constituents, including the public.

The cities of Boston, Denver, and San Antonio and the groups who conducted the evaluations are known to have presented the evaluation results to oversight bodies (Superintendent and Boston Public Schools Committee, Board of Education, and the Denver Preschool Program Board, respectively). Edvance, the organization that conducted San Antonio's evaluation, also presented results of the evaluation to the program's center directors. And NIEER's Preschool Matters newsletter included articles about the evaluations in both New Jersey and Oklahoma.

Results of the pre-k evaluations in the three States and three Cities were available to the public through either department of education websites and/or the websites of the organization that completed the evaluation. For example, Georgia and New Jersey posted evaluation reports on both the department of education websites and the website of external evaluators. Results of the Boston, Denver, and San Antonio evaluations were available on the respective department of education websites, while Oklahoma's evaluations were available on the websites of the external evaluators. Generally, evaluation reports were dated (and presumably posted) at the time of the release of the findings as outlined in Table 4. All of the evaluations received media coverage soon after the release of the findings, with most of the media coverage concentrated in the specific state or city (see Table 4). Many of the States and Cities used official press releases to garner initial media attention of positive evaluation findings.

New Jersey and Boston are two examples of how a State and a City used information gathered through program evaluation to improve their respective preschool programs. New Jersey began by collecting data on program implementation, classroom quality, and child outcomes, during the initial years of the Abbott preschool program. This information was used to focus statewide professional development and technical assistance to key areas that needed improvement. In 2002, New Jersey's Department of Education also formed the Early Learning Improvement Consortium, which consisted of the top early childhood faculty in the state, to monitor program improvement (Frede, 2005). Indeed, program quality did improve over the first decade of the Abbott preschool program (Frede, Jung, Barnett, & Figueras, 2009). New Jersey also has implemented a continuous improvement cycle whereby they use data to determine how to make improvements to better meet program standards (Frede, 2005). Like New Jersey, Boston also collects information on classroom quality, doing so every two years. Based on feedback from these observations, coaches work with teachers to build on their strengths and improve on their weaknesses (Shaw, 2014). Further, after a 2007 Boston Globe report regarding the mediocre quality of Boston's program, leadership overhauled the program by introducing evidence-based

curricula–Opening the World of Learning and Real Math Building Blocks: PreK–coupled with extensive coaching surrounding their implementation (Sachs & Weiland, 2010; Shaw, 2014).

Results of evaluations have also been used to promote preschool expansion both within the State or City as well as in other municipalities. As a result of their positive evaluation results, Boston expanded their preschool program in order to serve more children. They extended their public school-based model to include community-based partners, using lessons learned from the evaluation to improve implementation and partnering with Thrive in Five (Boston Public Schools Department of Early Childhood, 2015; Boston Public Schools, 2012; Yoshikawa, 2014). In New Jersey, the results of the evaluations have also been used to support the state's efforts to expand the targeted program throughout the state (Mead, 2009; Rundquist, 2013), both through the failed School Funding Reform Act of 2008 (Friedman et al., 2009) and through the recently awarded federal Preschool Development Grant (U.S. Department of Education, 2014).

Based on the strong results from their rigorous evaluations, Boston, Oklahoma, and New Jersey are frequently touted as exemplar programs for other states and cities seeking to begin, expand, or improve the quality of their preschool programs (Boston Public Schools, 2014; Graham, 2013; Mead, 2009; Shaw, 2014). For example, the book "Restoring Opportunity" features the Boston Pre-K program as a model program to reduce the school readiness gap, citing the results of the city's evaluation (Duncan & Murnane, 2014).

Table 4. Communication Strategy Matrix for Selected State and City Programs.

CHARACTERISTICS	STATES			CITIES		
	Georgia Bright from the Start	**New Jersey New Jersey's Abbott Preschool Program**	**Oklahoma Oklahoma's Early Childhood Program**	**Boston K0-K1**	**Denver Denver Preschool Program**	**San Antonio PrekSA**
Methods of communication regarding evaluation findings	No information about explicit communication strategies publicly available. Information about the evaluation included in the Decal newsletter in April 2014.[a]	Dissemination of evaluation findings through state website, media, and the Statewide Early Learning Coalition (Early Learning Improvement Consortium). Reports are released annually. The evaluation was discussed in a 2007 edition of NIEER's *Preschool Matters*.[d]	No information about explicit communication strategies publicly available. The evaluation was discussed in a 2006 edition of NIEER's *Preschool Matters*.[i]	Results of the evaluation were discussed during a briefing to the Superintendent of the Boston Public Schools (BPS) and the BPS School Committee in April 2012.[m]	Results of the evaluation were presented to the Denver Preschool Program (DPP) board (which has oversight for the program). t There is also an operations reports available via DPP website.[u]	Edvance shared findings with the Board of Education, school superintendents, the Mayor, the City Council, the City Manager, center directors, staff, and parents during the fall of 2014. The Communication Staff from the City of San Antonio also issued a press release for local media in late August 2014. The City of San Antonio uses a public relations/marketing contractor to garner media

						interest and a Communications Director to help with public relations and keeping the public informed.[x]
When did the program begin producing positive results that were causally linked to the program?						
When were evaluation results first published?	March 2014.[b]	December 2005, June 2007.[e]	November 2004, December 2005, June 2007.[j]	April 2012.[m]	November 2010.[u]	September 2014.[y]
Where were evaluation results first published?	Georgia Department of Early Care and Learning and Frank Porter Graham Child Development Institute Websites.[b]	New Jersey Department of Education – Division of Early Childhood Education and the National Institute for Early Education Research Websites.[e]	Georgetown University's Center for Research on Children in the U.S. and the National Institute for Early Education Research Websites.[j]	Boston Public Schools – Department of Early Childhood Website.[m]	Denver Preschool Program Website.[u]	Pre-K 4 San Antonio Website, Facebook page, and Twitter.[y]

	Georgia Bright from the Start	New Jersey New Jersey's Abbott Preschool Program	Oklahoma Oklahoma's Early Childhood Program	Boston K0-K1	Denver Denver Preschool Program	San Antonio PrekSA
Media Coverage of Evaluation	Beginning March 2014 with a press release from Georgia's Department of Early Care and Learning.[c] Limited media coverage found	Media coverage beginning December 2005.[f]	Media coverage beginning November 2004.[k]	Media coverage beginning October 2011.[n] Featured in Restoring Opportunity.[o]	Media coverage beginning February 2004.[v]	Media coverage beginning August 2014. City of San Antonio prepared a press release for local media. The City of San Antonio reports that the back-to-school time of year is ideal for releasing findings as there is heightened interest. The Program Evaluation Report was covered by Education Week, the Economist, and Univision.[z]
How is the evaluation used to help the program?	No information found.	New Jersey uses a Continuous Improvement Cycle. They also have an Early Learning	No information found.	Boston is expanding BPK including to community-based	DPP quality improvement process.[w]	In response to low CLASS scores in the Instructional

		Improvement Consortium. New Jersey collected information on classroom quality, implementation, and child outcomes at "baseline" which the state used to inform professional development and technical assistance. Early concerns about teacher coaches, use of time, and financial management were investigated and corrected.[g]		organizations. [p] The city implemented OWLS and Building Blocks (two research-proven curricula) with coaching to improve program quality after early reports of mediocre quality. Every 2 years an independent group monitors the program quality and provides feedback to teachers and coaches about their own classroom. The city has also encouraged NAEYC accreditation. [q] K-3 reform has aimed at sustaining BPK gains and aligning the curriculum with the Common Core.[r]		Support domain, Prek 4 SA coaches, center directors, and administrators were trained in the CLASS and now meet with teachers and teacher assistants to discuss quality improvement strategies. Prek 4 SA also has a strategic plan that is used as a continuous improvement cycle. Progress towards simplementatio n goals is reported on a monthly basis at Board meetings. Student assessments are conducted at the beginning, middle, and end of the year and are reviewed,

						along with classroom observations, to monitor classrooms and teachers.[aa]
	Georgia Bright from the Start	**New Jersey New Jersey's Abbott Preschool Program**	**Oklahoma Oklahoma's Early Childhood Program**	**Boston K0-K1**	**Denver Denver Preschool Program**	**San Antonio PrekSA**
How is the evaluation used to support preschool more generally?	No information found.	The positive results of the evaluation have been used to support preschool expansion in NJ and in other states and cities. [h]	Based on strong evaluation results, Oklahoma has been used as a model program to support preschool expansion in other states and cities. [l]	Based on strong evaluation results, Boston has been used as a model program to support preschool expansion in other states and cities. [s]	No information found.	Prek 4 SA has worked with other cities in Texas and across the country. [aa]

REFERENCES

a Cagle, B. (2014, April). Commissioner's Corner. *Decidedly DECAL.* Retrieved from http://www.decal.ga.gov/documents/attachments/DECALApril2014.pdf

b Bright from the Start (2015). *Evaluation of Georgia's Pre-K Program.* http://decal.ga.gov/BftS/EvaluationGAPreKProgram.aspx; Peisner-Feinberg, E. S., Schaaf, J. M., LaForett, D. R., Hildebrandt, L. M., & Sideris, J. (2014). Effects of Georgia's pre-k program on children's school readiness skills: Findings from the 2012-2013 evaluation study. Chapel Hill, NC: FPG Child Development Institute, The University of North Carolina at Chapel Hill. Retrieved from: http://fpg.unc.edu/sites/fpg.unc.edu/files/resources/reports-and-policy-briefs/GAPreKEval_RDDReport%203-4-2014.pdf

c Bright from the Start (2014, March 6). *Research Concludes Georgia's Pre-K Program Significantly Improves School Readiness Across Key Domains of Learning.* Retrieved from http://decal.ga.gov/documents/attachments/release_prekstudyphasetwo030614.pdf; Covert, B. (2014, March 10). Georgia's universal preschool program significantly improves children's skills. *Think Progress.* Retrieved from http://thinkprogress.org/education/2014/03/10/3385561/georgias-universal-preschool-scores/

d National Institute for Early Education Research (2007). Study of Abbott preschool program finds positive effects. *Preschool*

Matters. Retrieved from: http://nieer.org/publications/study-abbott-preschool-program-finds-positive-effects; Personal Communication W. Steven Barnett June 2015

e Barnett, W. S., Lamy, C., & Jung, K. (2005). *The Effects of Prekindergarten Programs on Young Children's School Readiness in Five States.* New Brunswick, NJ: National Institute for Early Education Research. Retrieved from http://www.nieer.org/resources/research/multistate/fullreport.pdf; Frede, E., Jung, K., Barnett, W. S., Lamy, C. E., & Figueras, A. (2007). *The Abbott Preschool Program Longitudinal Effects Study (APPLES): Interim Report*. New Brunswick, NJ: National Institute for Early Education Research. Retrieved from http://nieer.org/resources/research/APPLES.pdf; State of New Jersey Department of Education (1996-2010). *Early Childhood Education: Research*. Retrieved from: http://www.nj.gov/education/ece/research/studies.htm; Wong, V. C., Cook, T. D., Barnett, W. S., & Jung, K. (2007). An effectiveness-based evaluation of five state pre-kindergarten programs using regression-discontinuity. New Brunswick, NJ: National Institute for Early Education Research. Retrieved from http://nieer.org/resources/research/EvaluationFiveStates.pdf

f See NIEER "news releases" (http://nieer.org/news-events/news-releases) and "early education in the news" (http://nieer.org/news-events/early-education-news)

g Cerf, C. (2013). Self-Assessment Validation System (SAVS) for former Abbott Preschool Programs-Revised 2014-2015. New Jersey Department of Education, Division of Early Childhood Education. Retrieved from http://www.nj.gov/education/ece/savs/SAVSFormerAbbott1415.pdf; Frede, E. (2005). *Assessment in a Continuous Improvement Cycle: New Jersey's Abbott Preschool Program.* New Brunswick, NJ: National Institute for Early Education Research. Retrieved from http://nieer.org/resources/research/NJAccountability.pdf; Personal Communication W. Steven Barnett June 4, 2015

h Rundquist, J. (2013, March 20). Study: Preschoolers in poorest districts made 'significant gains' through 4th and 5th grades. *The Star Ledger*. Retrieved from http://www.nj.com/news/index.ssf/2013/03/nj_preschool_study_shows_gains.html; Mead, S. (2009). Education reform starts early: Lessons from New Jersey's PreK-3rd reform efforts. *Education Policy Program*. New America Foundation. Retrieved from http://fcd-us.org/sites/default/files/Education%20Reform%20Starts%20Early%20New%20Jersey.pdf

i National Institute for Early Education Research (2006). NIEER multi-state study finds meaningful gains from pre-K. *Preschool Matters.* Retrieved from http://nieer.org/publications/decemberjanuary-2006-%E2%80%93-volume-4-number-1

j Barnett, W. S., Lamy, C., & Jung, K. (2005). *The Effects of Prekindergarten Programs on Young Children's School Readiness in Five States.* New Brunswick, NJ: National Institute for Early Education Research. Retrieved from http://www.nieer.org/resources/research/multistate/fullreport.pdf; Gormley, W., Gayer, T., Phillips, D., Dawson, B. (2004). *The Effects of Oklahoma's Universal Pre-Kindergarten Program on School Readiness: An Executive Summary*. Center for Research on Children in the U.S. Retrieved from: https://georgetown.app.box.com/s/hxy0bp4dr3xrjyuqbimi; Wong, V. C., Cook, T. D., Barnett, W. S., & Jung, K. (2007). An effectiveness-based evaluation of five state pre-kindergarten programs using regression discontinuity. New Brunswick, NJ: National Institute for Early Education Research. Retrieved from

http://nieer.org/resources/research/EvaluationFiveStates.pdf

k Oklahoman (2004, November 21). Early start: Study hails Tulsa pre-K success. *NewsOK*. Retrieved from http://newsok.com/early-startbrstudy-hails-tulsa-pre-k-success/article/2874877; Toppo, G. (2008, June 26). Universal preschool students perform better. *USA Today.* Retrieved from http://usatoday30.usatoday.com/news/education/2008-06-26-preschool-universal_N.htm

l Graham, G. (2013, March 11). Tulsa's preschool programs seen as national model. *Tulsa World*. Retrieved from: http://www.tulsaworld.com/news/local/tulsa-s-preschool-programs-seen-as-national-model/article_a932f79f-ec5c-5619-a89a-0ef88d271830.html; Rowe, C. (2014, September 20). Free pre-k: Strong early gains, but long-term questions. *The Seattle Times.* Retrieved from http://www.seattletimes.com/education/free-pre-k-strong-early-gains-but-long-term-questions/; Khim, S. (2013, February 14). Is Oklahoma the right model for universal pre-k? *The Washington Post.* Retrieved from http://www.washingtonpost.com/blogs/wonkblog/wp/2013/02/14/is-oklahoma-the-right-model-for-universal-pre-k/

m Boston Public Schools (2012). Early Childhood Update: Presentation to school committee. Boston, MA. Retrieved from http://bpsearlychildhood.weebly.com/uploads/1/0/1/3/10131776/school_committee_4_11_12__early_childhood_presentation_3.pdf; Weiland, C., & Yoshikawa, H. (2012). Impacts of BPS K1 on children's early numeracy, language, literacy, executive functioning, and emotional development. Presentation to the School Committee, Boston Public Schools. Retrieved from http://bpsearlychildhood.weebly.com/uploads/1/0/1/3/10131776/2012_hgse_k1_impact_study.pdf

n Kristof, N. (2011, October 19). Occupy the classroom. *The New York Times.* Retrieved from http://www.nytimes.com/2011/10/20/opinion/occupy-the-classroom.html

o Duncan, G. J., & Murnane, R. J. (2013). *Restoring Opportunity: The Crisis of Inequality and the Challenge for American Education*. Harvard Education Press: Cambridge, MA.

p Boston Public Schools (2012). Early Childhood Update: Presentation to school committee. Boston, MA. Retrieved from http://bpsearlychildhood.weebly.com/uploads/1/0/1/3/10131776/school_committee_4_11_12__early_childhood_presentation_3.pdf; Yoshikawa, H. (2014). Investing in our future: The evidence base on quality preschool education. Symposium on Children. Crane Center for Early Childhood Research and Policy, Ohio State University; United Way of Massachusetts Bay and Merrimack Valley (2013). Press Release: Mayor Menino and Superintendent Johnson announce expansion of Kindergarten programs for city children. Retrieved from http://supportunitedway.org/press/2013/01/02/mayor-menino-and-superintendent-johnson-announce-expansion-kindergarten-programs-ci

q Sachs, J., & Weiland, C. (2010). Boston's rapid expansion of public school-based preschool: Promoting quality, lessons learned. *Young Children, 74-77.*; Shaw, L. (2014, January 30). How Boston's preschools went from mediocre to outstanding. *The Seattle Times.* Retrieved from http://blogs.seattletimes.com/educationlab/2014/01/30/how-bostons-preschools-went-from-mediocre-to-outstanding/; Yoshikawa, H. (2014). Investing in our future: The evidence base on quality preschool education. Symposium on Children. Crane Center for Early Childhood Research and Policy, Ohio State University.

r Boston Public Schools (2012). Early Childhood Update: Presentation to school committee. Boston, MA. Retrieved from http://bpsearlychildhood.weebly.com/uploads/1/0/1/3/10131776/school_committee_4_11_12__early_childhood_present

ation_3.pdf; Yoshikawa, H. (2014). Investing in our future: The evidence base on quality preschool education. Symposium on Children. Crane Center for Early Childhood Research and Policy, Ohio State University.

s Shaw, L. (2014, January 30). How Boston's preschools went from mediocre to outstanding. *The Seattle Times.* Retrieved from http://blogs.seattletimes.com/educationlab/2014/01/30/how-bostons-preschools-went-from-mediocre-to-outstanding/;

t Personal communication from APA consulting June 3, 2015

u Denver Preschool Program (no date). Research & Results: Our Results. Retrieved from: http://www.dpp.org/results-and-research/our-results

v The Denver Post (2001, February 9). Denver Preschool Program scores well in evaluation. *The Denver Post.* Retrieved from http://www.denverpost.com/breakingnews/ci_17340279; The Denver Post Editorial Board (2013, August 9). Denver preschool program making a difference. *The Denver Post.* Retrieved from: http://www.denverpost.com/editorials/ci_23831558/denver-preschool-program-making-difference; Gottlieb, A., & Pinney, T. (2004, February 1). Early childhood programs critical, studies prove. *The Denver Post.* Retrieved from http://nieer.org/news-events/early-education-news/early-childhood-programs-critical-studies-prove

w Robertson, N., McClelland, A., Palaich, B., Rooney, K., & Workman, S. (2015). Long term persistence effects of the Denver Preschool Program. In E. H. Greenberg (Chair), *New Evidence on the Longer-Term Impacts of State Preschool Programs*. Paper Symposium at the Association for Education Finance and Policy, Washington D.C.

x Personal communication from Edvance June 9, 2015

y Pre-K 4 San Antonio (2014). *Student Assessment.* Retrieved from http://www.sanantonio.gov/Pre-K4SanAntonio/Facts/StudentAssessment.aspx; Pre-K 4 San Antonio (2014). *Pre-K 4 SA Student Assessment Results for 2013-2014.* Retrieved from http://www.sanantonio.gov/Portals/0/Files/PreK4SA/ProgramDetails/StudentAssessmentResults2013-2014.pdf

z Aldridge, J. (2014, August 6). Pre-K 4 SA earns high marks for student achievement. *San Antonio Business Journal.* Retrieved from http://www.bizjournals.com/sanantonio/news/2014/08/06/pre-k-4-sa-earns-high-marks-for-student.html; Ehnstrom, J. (2014, September 9). City: Pre-K 4 SA program returns positive first-year results. *Ksat news.* Retrieved from http://www.ksat.com/content/pns/ksat/news/2014/09/09/san-antonio_s-pre-k-program-returns-positive-first-year-results.html; Pre-K 4 San Antonio (2014, August 6). Assessment shows Pre-K 4 SA students improved their developmental skills. Retrieved from http://www.sanantonio.gov/Pre-K4SanAntonio/pressrelease/TabId/936/ArtMID/2752/ArticleID/1899/Assessment-Shows-Pre-K-4-SA-Students-Improved-Their-Developmental-Skills.aspx; Pre-K 4 San Antonio (2014, September 9). Inaugural year of Pre-K 4 SA shows promising results. Retrieved from http://www.sanantonio.gov/Pre-K4SanAntonio/pressrelease/TabId/936/ArtMID/2752/ArticleID/1979/Inaugural-Year-of-Pre-K-4-SA-Shows-Promising-Results.aspx; Pre-K 4 San Antonio (2015, January 29). Pre-K 4 SA to enroll 1,700 students for 2015-16 school year. Retrieved from http://www.sanantonio.gov/Pre-K4SanAntonio/pressrelease/TabId/936/ArtMID/2752/ArticleID/2390/Pre-K-4-SA-to-

Enroll-1700-Students-for-2015-16-School-Year.aspx; Thomas, M. W. (2014, September 11). Pre-K 4 SA program received high marks in study. *San Antonio Business Journal.* Retrieved from http://www.bizjournals.com/sanantonio/blog/morning-edition/2014/09/pre-k-4-sa-program-receives-high-marks-in-study.html;

aa Personal communication from the City of San Antonio (Pre-K 4 San Antonio office) June 9, 2015

3.3. Comparison of preschool classroom measures for quality of care

The examination of quality in preschool evaluations is an important component of understanding what is determining and explaining results found in children. This requires examining the preschool classrooms with established instruments to measure the quality of the classrooms along various dimensions. This provides a clear report of the status of the classrooms across the city/state and can be conducted by sampling classrooms rather than examining every classroom in the state. Analyses of power can be performed to determine the number of classrooms that would need to be observed and evaluated to provide a reasonably precise and representative sample. Examinations of quality allow answering questions such as (1) How does the quality of preschool classrooms differ across providers? And/or (2) What is the impact of quality of the preschool experience on student outcomes? They require an observation of preschool classrooms for quality by an outside observer, and might require a potential increase of sample size in the evaluation. Having these as part of an evaluation provides states a look at program improvement over time and allow controlling estimates on child outcomes for levels of quality in programs, which increases precision of estimates. We described the most-used instruments in the field, together with some less frequently used instruments that target either specific populations or domains in the following sections. Table 5 summarizes the focus of these various observation tools, and whether validity and reliability have been established.

Table 5. Summary of Instruments for Examining Classroom Quality.

Instruments	Focus	Protocol	Validity	Reliability
ECERS-R	Overall program quality	Observation	Yes	Yes
ECERS-3	Overall program quality	Observation	Yes	Yes
CLASS Pre-K	Classroom practices & Interactions	Observation	Yes	Yes
SELA	Supports of children's literacy development	Observation	Yes	
SNAPSHOT	How children and teachers spend their time in the classroom	Observation	-	Yes
ELLCO	Language and literacy supports	Observation	Yes	Yes
CASEBA	Supports for dual language learners	Observation	-	Yes
PRISM	Supports mathematics and science learning	Observation	-	Yes
TSEEQ	Classroom practices and quality	Questionnaire	-	Yes

a) Frequently-Used 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, and a total scale score is averaged across all 43 items in the scale. This instrument provides an excellent look at the quality of classrooms in a program.

Early Childhood Environment Rating Scale - Third Edition (ECERS-3; Harms, Clifford, & Cryer, 2014). This is the newest observation tool on the field and consequently less used than the ECERS-R. Overall program quality is assessed by trained observers using this standardized measure of preschool classroom structure and process. This measure is the third edition of the ECERS, which 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 on the subscales, as measured by Cronbach's alpha, is reported by the authors to be adequate, ranging from .87 to .96 and .93 for the full scale. Classroom quality is rated on a 7-point Likert scale, indicating a range of quality from inadequate (1) to excellent (7). The six ECERS-3 subscales are as follows: Space and Furnishings, Personal Care Routines, Language-Reasoning, Learning Activities, Interaction, and Progam Structure. Average subscale scores are calculated, and a total scale score is averaged across all 35 items in the scale. This instrument provides an excellent look at the quality of classrooms in a program.

Classroom Assessment Scoring System (CLASS; Pianta, LaParo, & Hamre, 2008)
This is an observational system that assesses classroom practices from 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 CLASS provides information on 10 dimensions of quality that reach beyond the realm of environments and focus more specifically on teacher interactions and other features of instruction. The emotional support domain is measured through the use of four dimensions: Positive Climate, Negative Climate, Teacher Sensitivity, and Regard for Student Perspectives. The CLASS also measures Classroom Organization through three dimensions: Productivity, Behavior Management, and Instructional Learning Formats; and Instructional Support through three dimensions: Concept Development, Quality of Feedback, and Language Modeling. The CLASS can be used to reliably assess classroom quality for research and program evaluation and also provides a tool to help new and experienced teachers become more effective. CLASS is widely used in pre-K and K classrooms as it describes multiple dimensions of teaching that are linked to student achievement and development and has been validated in over 2,000 classrooms. 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). Large-scale research studies have shown that reliability of CLASS scores across observers, cycles, days and school years are strong and stable (Pianta, LaParo, & Hamre, 2008).

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.

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). Observations consist of time-sampled codes assigned to teacher and child behaviors, every 60 seconds (representing one cycle) over the course of the morning. Typically, four children are randomly selected from each classroom and each child is observed for 40 seconds, followed by 20 seconds of coding. This sequence is repeated for 2 to 3 hours in each classroom. Codes are divided into five subscales, including activity setting (i.e., whole group, free choice, transitions); peer interaction (simple social, cooperative pretend); child engagement (i.e., science, mathematics, oral language development); teacher-child engagement (i.e., scaffolds, didactic); and one-on-one teacher-child interactions (elaborated, routine).

The Early Language and Literacy Classroom Observation (ELLCO; Smith, Brady, & Anastasopoulos, 2008) is a comprehensive set of observation tools for measuring the level of classroom supports for language and literacy development. The ELLCO Pre-K comprises five sections: Classroom Structure, Curriculum, The Language Environment, Books and Book Reading, and Print and Early Writing. These five sections are grouped into two main subscales: the General Classroom Environment subscale, which consists of the Classroom Structure and Curriculum sections; and the Language and Literacy subscale, which comprises The Language Environment, Books and Book Reading, and Print and Early Writing sections. Reliability analyses have shown high Cronbach Alpha's ranging from .723 for the Curriculum section to .894 for the Print and Early Writing section. The tool has also shown good test-retest reliability, and sensitivity to interventions that target literacy.

Caregiver Interaction Scale (CIS, Arnett, 1989). The CIS has been widely used and measures teachers' behavior that focuses on teacher's interactions with children, their emotional tone, their discipline style and their responsiveness to children. The scale is composed of 26 items on a1 to 4 Likert scale, indicating the extent to which a teacher shows a particular behavior (1 = not at all, 2 = somewhat, 3 = quite a bit, and 4 = very much). The measure has shown four factors: teachers' sensitivity, harshness, detachment, and permissiveness. The sensitivity factor includes developmentally appropriate interactions, enthusiasm, and warmth. The harshness factor has to do with hostility and excessive critical behavior toward children. The detachment factor reflects the degree to which the caregiver is uninvolved or uninterested in children. The permissiveness factor refers to the tolerance of misbehavior.

b) Specialized/Other Instrumentation

The Classroom Assessment of Supports for Emergent Bilingual Acquisition (CASEBA; Freedson, Figueras & Frede, 2008). The CASEBA is designed to assess the degree to which preschool teachers and classrooms provide support for the social, cognitive and linguistic development of preschool-aged dual language learners (DLLs), with a focus on language and literacy. The instrument consists of 26 distinct rating scale items which cluster around six broad aspects of the early childhood curriculum: 1) teacher knowledge of child background information, 2) supports for home language and literacy development, 3) supports for English language and literacy development, 4) social-emotional supports and classroom management, 5) curriculum content, and 6) assessment. Each item is rated on a 7-point Likert scale, where 7 indicates that a specific form of support and accompanying practices are present in close to an ideal form, while 1

represents the total absence of any such practices. A modified version of this instrument can also be used to assess language and literacy teaching practices and supports for all children, regardless of home language. Though the CASEBA has not yet been publicly disseminated, a validity study of the measure was conducted in 100 classrooms in New Jersey, as well as in a staffing study of dual language learners in an urban district in New Jersey. Findings from the validity study are reported in *Dual Language Learners in the Early Childhood Classroom* (Freedson, Figueras-Daniel, Frede, Jung & Sideris, 2011). Although the tool is unpublished; it has garnered attention from early childhood dual language researchers across the country who are seeking observational tools to use in classrooms dominated by dual language learners, to assess the quality of teacher input and interactions for both research and professional development. Most recently, CASEBA was highlighted as a valuable tool to focus on the extent to which teacher practices and classroom quality specifically address the needs of DLL children (Castro, Espinosa & Paez, 2011).

Preschool Rating Instrument for Science and Mathematics (PRISM; Stevenson-Boyd, Brenneman, Frede, & Weber, 2009). The PRISM is a 16-item classroom observation instrument designed to measure the quality of materials and staff interactions to support preschoolers' mathematics and science learning. The 11 mathematics items in the PRISM provide a comprehensive picture of instructional supports for a wide range of mathematics skills and reasoning, including typically neglected areas such spatial reasoning, patterns, sequencing, and measurement, and its items and indicators are informed by the NAEYC/NCTM (2002). Two science items on the PRISM focus on materials that support explorations of biological and non-biological science and that encourage reading and writing about science. Three items focus on teaching interactions that encourage children to investigate, experiment, and discuss scientific concepts; support observing and predicting; and encourage children to record science information from their investigations. All PRISM items are scored on a seven-point scale with anchor points at the odd numbers: 1 = no materials or interaction or poor quality; 3 = minimal quality; 5 = good quality; and 7 = excellent quality. At each level, one or more indicators serve as evidence that a classroom has achieved that level of quality. The PRISM is currently being used in large-scale studies in New Mexico and New Jersey.

Teacher Survey of Early Education Quality (TSEEQ; Hallam, Rous, Riley-Ayers, & Epstein, 2011). The TSEEQ is a self-report survey for early childhood teachers regarding their classroom practices and quality. The survey is completed independently and can be conducted either on paper or online. Teachers are asked to reflect on several aspects of the curriculum and classroom practices. These areas include: literacy, math, science, physical education and art curriculum, curriculum in general, instruction, assessment, physical environment, interaction and emotional climate, leadership and supervision, and family involvement. There are approximately 100 questions on the survey, presented mostly with a 5-point Likert Scale response or a yes/no response. The complete survey is expected to take approximately 30 minutes per teacher. In a preliminary test of reliability, 490 surveys were analyzed and showed moderate to high levels of internal consistency, with low inter-item correlations (expected) and Cronbach alphas above .7.

c) Uses in State and City Preschool Systems

Table 6 below outlines the instruments used by various preschool systems, general uses, and frequency. While there is some variation across programs, ECERS-R has been the most widely used instrument. CLASS Pre-K has been used in various programs as well, but to a lesser extent. The new version of the ECERS, ECERS-Third edition is not included in this table because it is very recent and has yet to permeate

programs and QRIS systems. The dual use of quality of care measures as a program improvement tool and a quality tracking tool is quite frequent across most programs.

Table 6. State and City Preschool Programs and their Uses of Tools of Classroom Quality.

Pre-k Program	Measure							Uses		Frequency of observations
	ECERS-R	CLASS Pre-K	SELA	SNAPSHOT	ELLCO	PRISM	TSEEQ	Program Improvement	Track Quality	
STATES										
Alabama First Class Voluntary Pre-Kindergarten Program	X				X			X	X	1 x/year
Alaska Prekindergarten Program	X	X						X	X	1 x/year
Arkansas Better Chance	X							X	X	1 x/2 years
California State Preschool Program	X							X		1 x/year
Connecticut School Readiness	X								X	1 x/year
DC Public Charter School Pre-Kindergarten		X							X	Based on renewal or low performance
DC Public School Pre-Kindergarten		X						?	?	1 x/year
Georgia Pre-K Program		X						X	X	1 x/year
Illinois Preschool for All	X							?	?	?
Iowa Statewide Voluntary Preschool Program								X	X	Based on years in program
Kentucky Preschool Program	X							X	X	1 x/5 years
Louisiana Cecil J. Picard LA4 Early Childhood Program	X	X						X	X	As needed; no more than 1 x/year
Louisiana Non-Public Schools Early Childhood Development Program					X			X	X	2 x/year
Massachusetts Universal Pre-Kindergarten and Grant 391 Program	X								X	Based on self-assessed QRIS level
Michigan Great Start Readiness Program								X	X	3 x/year
Nebraska Early Childhood Education Program	X							X	X	Based on yr of pre-K grant
Nevada State Prekindergarten Education	X				X			X	X	1 x/year

Program									
New Jersey Former Abbott and Expansion Districts	X			X	X		X	X	1 x/year
New Mexico Pre-K	X						X	X	1 x/year
North Carolina NC Pre-K	X							X	1 x/3 years
Ohio Early Childhood Education							X	X	Based on QRIS step
Oregon Head Start Prekindergarten		X					X	X	1 x/3 years
Pennsylvania Pre-K Counts	X						X	X	1 x/2 years
Rhode Island Prekindergarten Program	X	X					X	X	Based on protocol
Vermont Early Education Initiative	X						X	X	1 x/3 years
Vermont Prekindergarten Education - Act 62	X						X	X	1 x/3 years
West Virginia Universal Pre-K	X	X					X		1 x/year
CITIES									
Boston's K0-K1	X	X		X			X	X	1 x/year
Denver's Preschool Program		X					X	X	?
Providence, Rhode Island Pre-K[a]	X	X			X		X	X	1 x/year
San Antonio - PreK 4 SA	X		X			X	X	X	1 x/year

Note: Oklahoma does not use classroom observations. Source: Expanded based on Ackerman, 2014. [a] Also used CASEBA.

d) Concerns and Issues

Concerns about measures of classroom quality have centered on two aspects, that is, concerns about the predictive validity of the measures, and concerns about the cultural appropriateness of the measures.

Concerns about predictive validity. A robust analysis on the predictive validity of classroom observation measures was carried out by Burchinal, Kainz and Cai (2011). The authors looked at how well our measures of classroom quality (discounting the most recent ECERS-3) predict child outcomes. The authors conducted a meta-analyses of studies that related such measures to child outcomes, and they also reanalyzed a set of large studies (the NICHD Study of Early Child Care and Youth Development; the Cost, Quality and Outcomes Study; The National Center for Early Development and Learning 11-State Evaluation; and the Head Start Family and Child Experiences Survey, known as FACES) that had information on low-income preschoolers and classroom quality. The authors found that higher quality programs (as measured through the CLASS, ECERS-R, CIS, and ORCE) were associated with higher language, academic, and social skills and lower behavioral problems. However, they also found that the associations are in fact quite modest (around .10 standard deviations) at most. Because high quality programs have shown large impacts on children, the fact that we are capturing such low associations is therefore interpreted by the authors as a problem with the quality measures themselves, and their lack of alignment with children's development, together with a lack of specificity.

Concerns about cultural appropriateness. While most of the measures summarized may include items on culturally responsive care, they may still be culturally disjointed. A first issue with cultural appropriateness has to do with understanding the norm against which observations tools were validated. This has been particularly criticized for the CLASS. Downer, Lopez, Grimm, Hamaguri, Pianta et al., (2012) argue that while the CLASS was not developed or normed specifically for DLLs, the structure and predictive validity of the CLASS applies equally well across preschool classrooms with different DLL compositions at the end of preschool. However, the underlying study was not meant to be representative of DLLs, nor was it meant to be a DLL study. A study on CLASS in distinctively different classrooms in terms of Hispanic composition shows a lack of generalizability of the CLASS across classrooms. The author argues this does not suggest that some aspects in the CLASS, such as providing opportunities for higher level thinking and maximizing learning opportunities, are not important for Hispanic children, but rather that the CLASS is an inappropriate assessment of teacher quality in these environments. A second issue with cultural appropriateness has to do with the relationship between quality and children's outcomes. Lopez, Arango, & Ferron, 2012 show that when DLL children are the primary focus, there is emerging data that suggests that there is no association between all subscales of the CLASS and child outcomes in Spanish and English. "The concern about the use of the CLASS for DLL children is that the indicators are not reflective of specific strategies needed to optimize child outcomes for DLL children." (Vitiello, 2013) These two issues in terms of cultural appropriateness have been less present in conversation about other quality observation measures. For the ECERS-R, Burchinal and Cryer (2004) showed that within the cultural variations found in the U.S., ECERS was a good predictor of child outcomes (Bryant, 2010).

3.3.2 Comparison of frequently used child assessment measures

The City of Seattle also requested a broad comparison of all widely available pre-K classroom and student assessments including any concerns about reliability, validity, or cultural appropriateness expressed by communities or the literature. As the number of assessments is quite extensive, beyond the scope of this work, we have focused instead on looking at assessments used in state and city program evaluations, assessments used by Early Achievers, and other frequently used assessment in preschool. In this section, we compare these, we then explore the different approaches to assessment: formative and summative and observer, teacher reported, or parent reported. This section finishes with general issues related to child assessment.

a) Assessing Children

In this section we are focusing on a subset instruments that have been used among the group of evaluations discussed in Table 3, in addition to measures currently being used in other types of preschool evaluation studies. For the most part, the most frequently used tests are the PPVT-III (although PPVT-V is now on the market and just now being incorporated into some evaluation work, at least at NIEER) and the Woodcock Johnson applied problems measure for math, in addition to other measures of vocabulary or language from the Woodcock Johnson. We believe this is partly due to the ease of use for these tools, as well as their availability in Spanish and English, the fact that these have actually been normed in both languages, and that they are measures that are collected by external individuals, rather than based on parent reporting.

Table 7 characterizes, for the group of selected measures, the target age, the domain the assessment is focused on, the language(s) in which the assessment is available, strengths, weaknesses, the purpose of the instrument, the normative data for it, and the reliability and validity information.

Table 7. Instruments for Examining Children of Preschool Age

Assessment (Author); Source(s)	Ages	Assessment Domain	Language	Strengths	Weaknesses	Purpose of Assessment (summative/ formative)	Purpose of assessment (clinical/ screening)	Normative data	Reliability (Consistent results regardless of assessor)	Validity (Depicts results on target outcome)
ASSESSOR REPORT										
PPVT III (Peabody Picture Vocabulary Test - Third edition) (Lloyd M. Dunn, PhD, Douglas M. Dunn, PhD/ Pearson); [B]	Age 2.5 and above.	Receptive Vocabulary	TVIP available in Spanish	Has been used widely in the developing world. Easy to administer. Picture based so no need for extensive translations. [C]	Some words or concepts may not be culturally appropriate. Children without experience decoding pictures will lose points. [C]	Summative	Assess vocabulary development from preschool through adulthood	Has reference norms in English and Spanish. Major limitation to Spanish norms is that the sample was small and homogenously high SES. [C]	Internal Consistency: Split-half correlation based on all subjects in the standardization sample were obtained. The coefficients on Form L ranged from .80 to .83. (only Form L was administered to adults). [D]	Considerable amount of evidence for supporting the validity of the instrument. [D]
PPVT-IV (Peabody Picture Vocabulary Test - Fourth edition; Lloyd M.	2:6 - 90+ Years	Receptive Vocabulary	TVIP available in Spanish	Supplies two equivalent forms of	Very recent. This makes it	Summative	Assess vocabulary developm	Has reference norms in	Provides extremely reliable scores, with all reliability and	Provides extremely reliable scores, with all reliability and

Dunn, PhD, Douglas M. Dunn, PhD/ Pearson)				the test with different vocabulary items to ensure children have not "learned the test"	pretty new to the evaluation field.		ent from preschool through adulthood	English. Spanish version is still the TVIP.	validity coefficients in the .90s range.	validity coefficients in the .90s range.
Woodcock-Johnson (Riverside Publishing); [E]	Age 2 and above	Intelligence Test b	English and Spanish	Comprehensive assessment of aptitude and achievement. Shown sensitivity to many types of interventions. Can be used across wide age range. [C]	Expensive ($1000 for initial kit). Needs large amount of equipment. Some concerns regarding comprehensiveness, appropriateness, and sensitivity. [F]	Summative	Used for research, to measure achievement, cognitive abilities, and to diagnose children's needs	Nationally representative sample of 8,818 subjects drawn from 100 US communities using stratified random sampling. Also normed in Spanish. [C]	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). Considerable amount of evidence for supporting the validity of the instrument. A

Assessment (Author); Source(s)	Ages	Assessment Domain	Language	Strengths	Weaknesses	Purpose of Assessment (summative/ formative)	Purpose of assessment (clinical/ screening)	Normative data	Reliability (Consistent results regardless of assessor)	Validity (Depicts results on target outcome)
Pre-CTOPPP (PreSchool Comprehensive Test of Phonological and Print Processing; Christopher J. Lonigan, Richard K. Wagner, Joseph K. Torgesen, & Carol A. Rashotte) M Note: The TOPEL is the published version of this measure in English. There is no published Spanish version.	3 to 5	Early literacy: phonological sensitivity, phonological memory, and phonological access. Includes a Print Awareness subtest (described below) and a Reading Vocabulary subtest.	English and Spanish	Greater depth as a literacy measure.	The Spanish version is a direct translation of the English version.	Summative	Research, screening	The normative sample consists of 842 preschool-aged children (3 to 5 years), residing in 12 states	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.	Content: Research indicated the areas assessed in PreCTOPPP provided unique aspects of early literacy important to predicting later reading skills. More specific information is provided under the published version TOPEL.
TOPEL (Test of Preschool Early Literacy; Lonigan,	3 to 5	Early literacy: knowled	English	Greater depth as a literacy	Hispanic-American Bilingual	Summative	Research, screening	The normative	The reliability evidence for the subtests and the	Content: Research indicated the 3 areas assessed in

C., Wagner, R., Torgesen, J., & Rashotte, C.) [G H]	ge of vocabulary, phonological awareness, and print knowledge	measure.	homes performed below average in validation.		sample consists of 842 preschool-aged children (3 to 5 years), residing in 12 states	composite score are good and range from .87-.96 for internal consistency, .81-.91 for test-retest, and .96-.98 for inter-scorer differences.	TOPEL provided unique aspects of early literacy important to predicting later reading skills. Criterion Prediction Validity: Correlations between TOPEL means and standard deviations with TERA were mostly large or very large. Construct Identification Validity: Age differentiation, gender and ethnicity, and language skills were examined. Differential Item Functioning: TOPEL mean scores were within average range for all groups, but Hispanic-American Bilingual homes

	performed below average.

Assessment (Author); Source(s)	Ages	Assessment Domain	Language	Strengths	Weaknesses	Purpose of Assessment (summative/ formative)	Purpose of assessment (clinical/ screening)	Normative data	Reliability (Consistent results regardless of assessor)	Validity (Depicts results on target outcome)
ELSA (Early Literacy Skills Assessment; Andrea Debruin-Parecki/Highscope) [I J K]	Pre-Kindergarten	Early literacy skills	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. Also, has two	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. [A]	Scores can be used for summative purposes or for tailoring early childhood programs to meet children's needs	n/a	English version normed with English speaking sample (over 500 for each version). Spanish version normed with a sample for which 80% of children had no or little understanding of English (307	Reliability estimates range from .6 to .8.	Correlates highly with WJ and TOPEL. The Spanish version was piloted with a variety of Spanish-speaking populations and also found to be reliable and valid [L]

				versions. [A]				children).		
PreLAS 2000 (De Avila & Duncan, 2000/McGraw Hill)[AD]	Pre-K-1	English Language Proficiency	English & Spanish	PreLAS helps measure language development of first– and second–language students in both English and Spanish	Developers have no reports on predictive validity or correlations with other language proficiency tests.	Summative	Research, screening/ placement	Standardized with sample of about 800 children; around 20% Spanish speaking.	Estimates at the subtest level range from .85 to .91. Test-retest coefficients for preLAS-2 subtests also adequate (around .90).	No reported relationship identified to other tests of language proficiency. Only criterion-related validity reported. Also, total test scores and proficiency levels found to distinguish children of English only backgrounds from those whose first language was not English.
EWA (Early Writing Assessment; Puranik & Lonigan, 2011) [AE]	3 - 5 year olds	Written language knowledge		English and Spanish	Lack of norms to compare to.	Summative	Research. Capture linear sequence in writing skills acquisition in preschoolers.	No norms.	Inter-rater reliability was calculated for the different tasks scored ranged from 93 to 100%. Internal consistencies were around .90.	No reported relationship identified to other tests.

Assessment (Author); Source(s)	Ages	Assessment Domain	Language	Strengths	Weaknesses	Purpose of Assessment (summative/ formative)	Purpose of assessment (clinical/ screening)	Normative data	Reliability (Consistent results regardless of assessor)	Validity (Depicts results on target outcome)
McCarthy Scales of Children's Abilities (MSCA) (Riverside Publishing)[N]	Ages 2 to 8	Comprehensive Measure: Cognitive and Motor.	English	Assessment of many developmental domains. Has been used widely in the developing world.[C]	Norms are outdated. Requires trained administrator.[C]	Scores can be used for diagnostic or summative purposes.	Research, screening.	Standardized on a sample of 1,032 children stratified by race, geographic region, father's occupational status, and urban-rural residency "Exceptional" children were excluded from the	Reliability coefficients for the general cognitive index tend to be around 0.90.	Correlations with the Stanford-Binet scale (Form L-M) and the WPPSI. The cognitive index correlates at .81 with the Binet IQ and at .71 with the WPPSI full-scale IQ.

								standardization sample.[C]		
Denver II (Denver Developmental Screening Test II) (Denver Developmental Materials) **NOTE: This instrument has been discontinued.**	Birth to 6	Fine motor/adaptive, gross motor, language, and personal/social.[C]	English and Spanish.	Assessment of some domains of development. Has been used widely in the developing world. Appropriate for children up to 6 years of age. [C]	Not designed to assess specifics of any particular construct (e.g. language). Does not yield continuous scores. Difficult to administer. [C]	Summative	Research. Clinical Screening: The DDST II does not provide continuous scores indicating children's developmental status, instead only providing an indication of whether the child appears to have developmental delays when compared to children of the same	Standardize with a sample of over 2,000 children, representing a broad spectrum, and representative of the Colorado population. Minor demographic differences between the sample and the U.S.[C]	Not available.	Studies have demonstrated that Denver II has good sensitivity but an unacceptably low specificity.

								age.[C]		
Assessment (Author); Source(s)	**Ages**	**Assessment Domain**	**Language**	**Strengths**	**Weaknesses**	**Purpose of Assessment (summative/ formative)**	**Purpose of assessment (clinical/ screening)**	**Normative data**	**Reliability (Consistent results regardless of assessor)**	**Validity (Depicts results on target outcome)**
Griffiths Mental Development Scales Extended revised (GMDSER) - 2006 [O]	2 - 8 years	Comprehensive	English	Covers a broad range of developmental domains; widely used, acceptable sensitivity and specificity, standardized recently on a UK population.	Intensive training required, no evidence for use as a population measure, no evidence about acceptability by parents, lengthy to administer, little published evidence on validity.	Summative	Research, screening	The measure was normed on a national representative sample of children in UK between 2-8 years of age.	With the exception of Scale E (performance) in children with age under 48 months the coefficients exceed 0.70.	Griffiths correlates well with the Bayley Scales of Infant Development and the Battelle Developmental Inventory for infants. But age equivalents obtained on the Griffiths were considerably higher than those obtained from the Bayley and the Battelle.
LENS (Lens on Science) [AF]	Pre-K	Science abilities.	English /tablet based.	Comprehensive Science	Expensive to administ	Summative	Research	No norms.	High person reliability (.93) and item	Predictable correlations with related measures,

	Measure.	er and somewhat lengthy.		reliability (.98) for the non-tablet version that predates the LENS.	growth in science ability across the preschool school year and sensitivity to detect the positive impact of a classroom based preschool science intervention for the non-tablet version that predates the LENS.

Assessment (Author); Source(s)	Ages	Assessment Domain	Language	Strengths	Weaknesses	Purpose of Assessment (summative/ formative)	Purpose of assessment (clinical/ screening)	Normative data	Reliability (Consistent results regardless of assessor)	Validity (Depicts results on target outcome)
Research-Based Elementary Mathematics Assessment [P]	3 to 6	Early Math	English	Based on learning trajectories. Assesses a wider range of early numeracy, geometry, and spatial skills	Recent adaptation of REMA used on the Boston evaluation only to our knowledge.	Summative	Research	Sample 1: 1,930 prekindergarten and kindergarten children. Sample 2: drawn from a randomized controlled trial of the Building Blocks curriculum in two large urban school districts in the north-	Boston study: Item reliability approximately of 1.00, suggesting that the difficulty rank order of this study's items would be constant or close to constant across different samples drawn from the population to which we generalize. Person reliability of .76. Within Sample 2, item reliability was also approximately 1.00, while the person reliability was .68. TEAM Validity: Item reliability was .98. Inter-rater reliability was .98	Short Form scores evinced strong correlations with full REMA scores within each time point: at the beginning of prekindergarten, .71 and at the end .74. The correlations between children's WJ Applied Problems and Short Form were .74.

								eastern USA beginning in the fall of 2006.		
BASC 2 (Behavior Assessment System for Children, Third Edition; Randy W. Kamphaus, PhD, Cecil R. Reynolds, PhD/Pearson) NOTE: BASC 3 to be released in 2015. [Q]	Ages 2 to 21	Social skills and behavior [B]	English, Spanish [B]	Teacher and Parent Ratings allows for flexibility. Very comprehensive.	The Spanish-speaking population is included in the standardization samples yet separate norms for this population are not available. It has limited validity information about applicability with preschoolers. [AB]	Summative	Research, screening and clinical. Useful in the classification of various emotional and behavioral disorders	General and clinical norm samples. The standardization sample closely matches the U.S. Census data with regard to gender, race/ethnicity, clinical or special education classification. [A]	The scales and composites have high internal consistency and test-retest reliability.	Construct validity, for the internalizing and externalizing dimensions of the BASC scales are supported by the results of a factor analyses and structural equation analysis. Criterion-related validity of the scales is satisfactory. Research supports the validity of the PRS and TRS for the assessment and identification of children presenting with attention-deficit/hyperactivity disorder. [R]

Assessment (Author); Source(s)	Ages	Assessment Domain	Language	Strengths	Weaknesses	Purpose of Assessment (summative/ formative)	Purpose of assessment (clinical/ screening)	Normative data	Reliability (Consistent results regardless of assessor)	Validity (Depicts results on target outcome)
Emotion Recognition Questionnaire (ERQ; Ribordy, Camras, Stefani, & Spaccarelli, 1988)		Social Skills: assesses children's ability to identify emotions	English	Simple to administer and score. There is also a nonverbal response adaption of the measure.	No adaption available for different cultures or language backgrounds.	Summative	Assess general development of aspects of behavioral regulation	These analyses were conducted on the first cohort on the high-risk control sample (n = 155) and the normative sample (n = 387, N = 463 with overlap) from the third year of the study	Four of the scales indicated a lower level of internal reliability for both the high-risk control sample and the normative sample. These scales were the Number Angry Correct, the Number Happy Correct, the Number Sad Correct, and the Total Number Correct. Higher Cronbach for Number Happy Correct.	The ERQ has been used with Head Start Children and demonstrated sensitivity to intervention effects [S]
TOQ (Smith-Donald, Raver et		Executive	English and	Simple to administer		Summative	Research	No norms.	Cronbach's alphas consistently above	The scale shown predictive validity

al. 2007)		function, compliance and attention	Spanish	. Validated in Spanish. Sensitive to maltreatment.					0.85.	of cognitive and socio-emotional outcomes and executive function measures and has been validated in the U.S. [T]
Head-Toes-Knees-Shoulders (HTKS; Ponitz etal., 2009) [v]	3 to 7	Executive Functions: inhibitory control and attention	English (NIEER translation available)	Quick and easy to administer and score. Performs strongly.	English version only.	Summative	Research	No norms.	HTKS .80	The Head Toes Knees Shoulder task has been widely used and is also a consistent predictor of emergent mathematics, vocabulary, and literacy [W]
Dimensional Change Card Sort (DCCS; Zelazo, 2006) [v]	ages 3-85	Executive Functions: attention shifting,	English (NIEER translation available)	Easy to administer.	English only.	Summative	Research	No norms.	DCCS .47 on the Vanderbilt study.	Predictive validity on the Vanderbilt study.
Peg Tapping (Diamond & Taylor, 1996) [v]	3-6	Executive Functions: cognitive inhibitory control	English (NIEER translation available)	Easy to administer.	English only.	Summative	Research	No norms.	Peg Tapping .80 on the Vanderbilt study.	Predictive validity on the Vanderbilt study; best performing measure.

Assessment (Author); Source(s)	Ages	Assessment Domain	Language	Strengths	Weaknesses	Purpose of Assessment (summative/ formative)	Purpose of assessment (clinical/ screening)	Normative data	Reliability (Consistent results regardless of assessor)	Validity (Depicts results on target outcome)
Copy Design [v]	3-6	Executive Functions	English (NIEER translation available)	Easy to administer.	English only.	Summative	Research	No norms.	Copy Design .72 on the Vanderbilt study.	Predictive validity on the Vanderbilt study.
TEACHER REPORT										
Teaching Strategies GOLD - Teaching Strategies, 2010	Birth through Kindergarten [A]	Comprehensive	English, Spanish [A]	Comprehensive; developmental trajectory of learning and development; gives insight for individualizing instruction [A]	May vary depending of teacher experience and training (teacher bias evident in one study); Large number of items to collect data on and evaluate; Minimal	Formative	Track child's efforts, achievements, and progress; designed to enhance instruction and improve learning	Nationally representative norm sample of 18,000 children from 50 states, PR, and DC; across age cohorts; Provides norm tables across	Strong internal consistency; Inter-rater reliability for kindergarten teachers seems weak.	Varying validity based on study report and age of children; seems to be most valid for math and literacy; low concurrent and construct validity with direct measures of language, self-regulation, and cognition.

		concurrent/ construct validity with direct assessments of language, cognition, and self-regulation. [A]	all six areas of development. Each norm table includes expected scores for children across 24 different 3-month age bands from 0-71 months. Includes norms for fall, winter, and spring.	

Assessment (Author); Source(s)	Ages	Assessment Domain	Language	Strengths	Weaknesses	Purpose of Assessment (summative/ formative)	Purpose of assessment (clinical/ screening)	Normative data	Reliability (Consistent results regardless of assessor)	Validity (Depicts results on target outcome)
Early Learning System (ELS) (Riley-Ayers, Stevenson-Garcia, Frede, and Brenneman 2012; Riley-Ayers, Stevenson-Garcia, Brenneman, Thompson, & Thompson, 2014/Lakeshore) [AA]	3-6 years [A]	Comprehensive	English	Measure that examines child development across several domains through a manageable assessment system for teachers. Provides valuable information for teachers to inform instruction. Provides a developmental trajectory of	Does not examine development of children in the arts; It is not used widely at this time. [A]	Formative [A]	Track child's efforts, achievements, and progress; designed to enhance instruction and improve learning. [A]	This assessment is not normed. [A]	High inter-rater reliability. [A]	Moderate correlations with other established tests. High internal consistency. [A]

			children's development. [A]					
Cooper-Farran Behavioral Rating Scales [AC]	Social skills and behavior	English	Quick and easy to complete. Teachers rate a total of 37 items. Uses 7 point scale and differentiates between children's interpersonal social skills and learning-related social skills.	Has been used primarily with children in K through 2nd grade and its accuracy with older children is unclear. Only teacher version available.	Summative	Assesses interpersonal and learning-related social skills for research or screening purposes.	No norms.	Intrarater reliability ranges from .49 to .80. Interrater reliability ranges from .31 to .68.

Assessment (Author); Source(s)	Ages	Assessment Domain	Language	Strengths	Weaknesses	Purpose of Assessment (summative/ formative)	Purpose of assessment (clinical/ screening)	Normative data	Reliability (Consistent results regardless of assessor)	Validity (Depicts results on target outcome)
High/Scope Child Observation Record (High Scope)	2 and a half to 6	Comprehensive	English and Spanish	Comprehensive; developmental trajectory of learning and development; gives insight for individualizing instruction	May vary depending on teacher experience and training; no normative data. No evidence of reliability or validity on its Spanish version.	Formative	Track child's efforts, achievements, and progress; designed to enhance instruction and improve learning	Not normed. Criterion referenced.	85.7% agreement for inter-rater reliability for 70 teachers; acceptable internal consistency.	Demonstrated significant differences in scores at age category and moderate to high correlations with other standardized measures.
PARENT REPORT										
Ages and Stages Questionnaire (ASQ) (Brookes Publishing);	1 month to 66 months	Comprehensive	English, Spanish, French, Korean	Covers developmental domains and produces scores for	Used primarily as screener. No evidence of	Summative	Comprehensive screening program, used to identify children	Standardized on 15,138 children from families of	Inter-observer agreement r=.92; Test-retest r=.95	The ASQ was validated against the Bayley Scales of Infant Development. Concurrent =.84; Sensitivity =.72;

	each, can be a population measure, flexibility to administer, gets parents involved, provides a good basis for discussion, can be used with children at high risk of developmental problems, quick and easy, one-off purchase	validity or reliability in Spanish.	who need additional evaluations	different educational and economic backgrounds	Specificity = .86. Evidence of capturing program effects.

Assessment (Author); Source(s)	Ages	Assessment Domain	Language	Strengths	Weaknesses	Purpose of Assessment (summative/ formative)	Purpose of assessment (clinical/ screening)	Normative data	Reliability (Consistent results regardless of assessor)	Validity (Depicts results on target outcome)
Ages and Stages Questionnaire (ASQ) : Socio Emotional (Brookes Publishing)	6 month to 66 months	Socio-Emotional	English and Spanish	Easy to administer, flexible and can be used to work with families and children.	Used primarily as screener. No evidence of validity or reliability in Spanish.	Summative	Comprehensive screening program, used to identify children who need additional evaluations	Developmental norms based on 3,014 preschool -age U .S. children.	Test-retest and internal consistency reliability.	Content and convergent validity. Evidence of capturing program effects. A sample of 1,041 children with completed ASQ-SE questionnaires were then assessed with either the Child Behavior Checklist (CBCL), the Vineland Social Emotional Early Childhood Scale (SEEC), or they had a professionally diagnosed social-emotional disability. The results of these screenings were then compared for this sample of children in order

										to determine appropriate cutoff points for the ASQ - SE.
Vineland Adaptive Behavior Scales, Second Edition (Pearson)	Birth to age 90-11	Communication, Daily Living Skills, Socialization, Motor Skills	English and Spanish	Easy to administer, can be done by sets, and has shown to differentiate well for cognitive delays and autism.	Lengthy set of question sets. Neither reliability nor validation of the Spanish version.	Summative	Screening and identification of developmental delays but must be combined with other assessment for diagnostic purposes	National sample of 3,695 stratified according to race/ ethnicity, geographical region, sex, and mother's education level	Internal consistency reliabilities .80 for the three primary domains	The Vineland-II domain scores tended to show moderately strong convergent correlations with comparable scales from the Adaptive Behavior Assessment System, Second Edition, with correlations averaging around .70 for similar X scales. With regard to discriminant validity, Vineland-II domain scores tended to correlate at rather low levels with intelligence test scores from the Wechsler tests, correlations generally falling in the range from .10 to .35

Assessment (Author); Source(s)	Ages	Assessment Domain	Language	Strengths	Weaknesses	Purpose of Assessment (summative/ formative)	Purpose of assessment (clinical/ screening)	Normative data	Reliability (Consistent results regardless of assessor)	Validity (Depicts results on target outcome)
SSRS (Social Skills Rating System; Teacher and Parent/Pearson) has been replaced by the SSIS (Social Skills Improvement System Rating Scales. [Y]	3:0 - 18:0	Social skills and behavior	English and Spanish	Flexible, allows using the teacher or parent versions.	No information on validity or reliability in Spanish.	Summative	Research, screening	National norms for pre-school	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.	Collection of reliability and validity evidence, including correlations with other measures and consistency with research on special populations
CBCL (Child Behavior Checklist) - or- C-TRF (Teacher Form) AG	1.5-5, 6-18	Children's emotional, behavioral and social aspects of life			Has a module with multicultural options. Low cost. Extensively used.	Summative	Research and screening/ clinical	Normed. Based on over 27,000 CBCLs and C-TRFs from 24 societies.		

A Reynolds, C. R., & Kamphaus, R. W. (2004). BASC-2: Behavior assessment system for children.

B Peabody Picture Vocabulary Test - Revised (PPVT-R). (2000). Retrieved June 10, 2015, from https://www.nlsinfo.org/content/cohorts/nlsy79-children/topical-guide/assessments/peabody-picture-vocabulary-test-revised

C Examining_ECD_Toolkit

D instdoc

E Woodcock-Johnson Test Guide. (2015). Retrieved June 10, 2015, from http://www.tests.com/Woodcock-Johnson-Testing
F Clements, D. H., Sarama, J. H., & Liu, X. H. (2008). Development of a measure of early mathematics achievement using the Rasch model: the Research-Based Early Maths Assessment. Educational Psychology, 28(4), 457-482.
G TOPEL: Test of Preschool Early Literacy. (2012). Retrieved June 10, 2015, from http://www.proedinc.com/customer/productView.aspx?ID=4020
H http://www.uofaweb.ualberta.ca/elementaryed/ccrl.cfm
I Authentic Research Based Field TestedEarly Literacy Skills Assessment (ELSA). (n.d.). Retrieved June 10, 2015, from http://www.highscope.org/file/Book-Pages/ELSAFullPresentationDocument0105.pdf
J DebRuin-Parecki, A. (2006). Early Literacy Assessment for Spanish-Speaking Preschoolers: Filling a Need. Retrieved June 10, 2015, from http://www.highscope.org/file/NewsandInformation/ReSourceReprints/Spring06pdfs/EarlyLiteracyAssess.pdf
K Cheadle, J. (2007, May 9). The Early Literacy Skills Assessment (ELSA): Psychometric Report for Both English and Spanish Versions. Retrieved June 10, 2015, from http://www.highscope.org/file/Assessment/ELSAJacobs.pdf
L DeBruin-Parecki, A. (2005b, Summer). The Early Literacy Skills Assessment in Spanish,
M Lonigan, C., Wagner, R., Torgesen, J., & Rashotte, C. (2002). Preschool Comprehensive Test of Phonological & Print Processing. Retrieved June 10, 2015, from https://psy.fsu.edu/~lonigan/01234Pre-CTOPPP.htm
N Kaplan, R. M., & Sacuzzo, D. P.(2010). Psychological Testing: Principles, Applications, & Issues, Eighth Edition. Belmont, CA: Wadsworth, Cengage Learning.
O Bedford, H., Walton, S., & Ahn, J. (2013, June 1). Measures of Child Development: A review. Retrieved June 10, 2015, from https://www.ucl.ac.uk/cpru/documents/review_of_measures_of_child_development
P (REMA; Clements et al., 2008) - SHORT FORM by Weiland et al., 2012.
Q BASC-2 Summary. (2015). Retrieved June 10, 2015, from http://basc-2.szapkiw.com/basc-summary/
R Doyle, A., Ostrander, R., Skare, S., Crosby, R. D., & August, G. J. (1997). Convergent and criterion-related validity of the behavior assessment system for children-parent rating scale. Journal of Clinical Child Psychology, 26(3), 276-284.
S (Bierman et al., 2008)
T Smith-Donald, R., Raver, C. C., Hayes, T., & Richardson, B. (2007). Preliminary construct and concurrent validity of the Preschool Self-regulation Assessment (PSRA) for field-based research. Early Childhood Research Quarterly, 22(2), 173-187.
V Lipsey, M., Nesbitt, K., Farran, D., Dong, N., Fuhs, M., & Wilson, S. (2014, May 1). Learning-Related Cognitive Self-Regulation Measures for Prekindergarten Children with Predictive Validity for Academic Achievement (Working Paper). Retrieved June 10, 2015, from https://my.vanderbilt.edu/cogselfregulation/files/2012/11/Self-Reg-summary-paper-5-7-141.pdf
W (Matthews, Ponitz, & Morrison, 2009; Turner et al.,2012).
X Children's Assessments (PPA Presentation) Page 7
Y PreK-16 Education and Special Needs. (2015). Retrieved June 10, 2015, from http://www.pearsonclinical.com/education/products/100000322/social-skills-improvement-system-ssis-rating-scales.html#tab-details
Z (HTKS; Ponitz etal., 2009)

AA (Riley-Ayers, Stevenson-Garcia, Frede, and Brenneman 2012; Riley-Ayers, Stevenson-Garcia, Brenneman, Thompson, & Thompson, 2014/Lakeshore)

AB Sandoval, J., & Echandia, A. (1995). Behavior assessment system for children. *Journal of School Psychology, 32*(4), 419-425.

AC Cooper, D., & Farran, D. C. (1991).

AD http://www2.uwstout.edu/content/lib/thesis/2003/2003sidersj.pdf

AE Puranik, C.S., and Lonigan, C.J. (2011). From Scribbles To Scrabble: Preschool Children's Developing Knowledge Of Written Language. *Reading and Writing, 24*(5): 567-589. Puranik, C.S. and Lonigan, C., (2012). Early Writing Deficits in Preschoolers with Oral Language Difficulties. *Journal of Learning Disabilities, 45*(2): 179-190.

AF https://www.sree.org/conferences/2011f/program/downloads/abstracts/392_4.pdf

AG http://www.aseba.org/preschool.html

b) Understanding the Various Approaches to Child Assessment

Formative versus Summative. Broadly speaking, the use of assessments can be described as formative or summative. Formative assessment is the use of assessment to inform teaching, with some definitions going so far as to equate formative assessment with scaffolding. Formative evaluation is internal and takes place during the educational experience. It looks forward in a process that is responsive to the needs of the learner. Summative assessment is the use of assessment to judge progress or attainment relative to a standard. Summative assessment of the performance of a child looks backwards and may be used to judge the contributions of a teacher or program to child progress. Summative assessment generally is external in its orientation. Summative assessments may be used to inform professional development and other supports for teachers and programs, but they also may be used to make "high stakes" decisions, including to sanction or reward teachers or schools, and to inform decisions about public programs and policies. In addition, summative assessments are commonly used to make high-stakes decisions about individual children, including the provision of additional supports (e.g., special education services and services for immigrant children who have limited proficiency in the local language) and opportunities (e.g., programs for gifted children), as well as to determine whether a child should enter primary school at the typical age or delay entry. As it is the use of an assessment that is formative or summative rather than the assessment instrument itself, the same instrument can be used for summative or formative purposes. Confusion can arise because instruments have been designed to be particularly useful for formative or summative purposes, and sometimes the instruments themselves are referred to as formative or summative measures.

Observer, Teacher, or Parent Reported. Information on children's learning and development can be collected through a variety of methods, both quantitative and qualitative. Information on children can be obtained directly from children or from those who observe them, most often parents and teachers or other adult caregivers.

External observers can provide accurate data on children, particularly when they have been trained to be reliable and accurate (with standards for interrater reliability ideally above .90). They are also less likely to have biases for particular children, since they lack a relationship with the children and their work will be based on children across a study, rather than in comparison to other children in the classroom. External observers do require careful training on child protocols for assessment, and have to be background checked and trained on the assessment itself. Some assessments do require extensive training and only external observers are likely to have the time for these, relative to teachers or parents.

Teachers in preschool or school settings often provide valuable insights into children's learning and development, though they can only report on those children who attend the program. Teachers make good informants because they tend to spend a great deal of time with the children, and have working knowledge of and/or training in learning and development. However, teachers vary considerably in their preparation and training. This can be expected to greatly affect their ability to evaluate children, especially with performance assessments. The less standardized and more qualitative an assessment, the more the quality of the results (validity, reliability, and fairness) depends on the teacher's knowledge and skills regarding children's learning and development, and the assessment itself or overall. For many instruments, specialized training of the teacher (or other assessor) may be required.

Parent/caregivers are valuable informants because of the intimate knowledge they acquire of a child, due to their relationship and the time they spend with the child. However, if caregivers are asked to provide ratings relative to an implicit standard or expectation (for example regarding learning,

development, relationship quality, life satisfaction, or happiness) they may differ greatly from one socio-economic environment and cultural background to another regarding what is typical or normative (Ertem et al., 2008). Caregivers also tend to provide socially desirable answers. Despite these disadvantages, caregivers' information about children can be valuable, and nationally representative information can be readily obtained through household surveys. Some checklists and rating scales are designed to be relatively robust with respect to variations among parents.

c) Concerns and Issues

Concerns about cultural appropriateness. Recommendations for the assessment of DLL children (Espinosa & Lopez, 2007) include examining children's vocabulary in the child's home language, as well as English, even without psychometrically sound approaches existing on how to combine these. They also recommend that results be interpreted in context, and take into account the child's early language experiences. Progress in both languages should be taken into account, so as to have a better sense of long-term outcomes, and as a consequence, assessments available in both languages should be chosen. In addition, Espinosa & Lopez recommend considering the proportion of DLL children when sampling for either program accountability, and/or research purposes, and to take into account the extent of home language supports in the children's preschool experiences. Research has also shown (Halle, et al., 2014) that socio-emotional outcomes may differ between DLLs and English speakers, and that the classroom's context and the use of the home language can moderate these results. As a consequence, analyses and interpretation of results even in the social-emotional dimension for DLL children should incorporate information on classroom supports for the home language, as well as children's language background.

Summary/conclusion

Our review provides a wealth of information that can be used to inform key decisions regarding the evaluation of Seattle pre-K. The most general conclusion is that a number of viable options are available for the overall research design. The choice of an overall design strategy probably is the most significant evaluation decision to be made. The particulars of data collection time, procedures, and instruments, are more closely circumscribed, as they are highly similar across the most successful programs and their evaluations. Fortunately, these details also are highly consistent with common current practices in Seattle and Washington. In addition, we believe that the review indicates that the success of the evaluation depends highly on the provision of effective supports to providers in the program. In some circumstances, a city might choose to delay evaluation for several years in order to assure that all processes are in place. However, this would mean that results were not available for some time, and some aspects of the evaluation will become more difficult as the program expands. If the decision is made to move ahead quickly, it is advisable to also plan for a second wave of evaluation when the program is more mature.

We reached no firm conclusion regarding choice of overall evaluation design. It is advisable to use more than one design strategy, as no strategy is without weaknesses, and the use of two or more designs greatly reduces the chances of a serious failure. For the first year of the study, with 14 classrooms, the following options are reasonable and have been adopted by one or more of the studies reviewed.

Option 1. Simply collect baseline data on program quality and child progress during the year. This information will establish a starting place for the study and serve as a comparison for the following year, regarding the program quality and characteristics of children enrolled. Use this information to inform

improvements to both the program and the research process in the following year. The disadvantage of this approach is that it can provide no serious estimates of program impact.

Option 2. Conduct a randomized control trial (RCT) through a lottery for program entry while encouraging oversubscription. The lottery will produce a waiting list group that can serve as a control group. If there is a sufficient pool of children to provide even 100 to 150 children on the wait list, then this approach has a strong chance of providing good estimates of program impact. Given the small scale of start-up, we also expect that the initial group of providers will be relatively strong, though it is important that this hold true. This approach provides impact results very quickly and sets up a strong longitudinal study. As the RCT children may be expected to disperse among kindergartens, the small sample size keeps cost reasonable in follow-up.

Option 3. Form a local comparison group using children who attend other programs. This comparison group could be obtained from a local sample that is already being assessed for other purposes, or by choosing specific providers in closely matched areas and assessing their children (we recognize that it will not be easy to obtain cooperation from such providers even if they are guaranteed anonymity). This allows for a pre-test for the comparison group. We do not recommend weaker options for Year 1.

We do not necessarily recommend an RDD study in Year 1, given the small sample size (though it might be tried purely for comparison to later years and the RCT). However, we do recommend that an RDD study be considered for Years 2 and 3 if it can be implemented in programs where the entering pool of children does not change from Years 2 to 3 to 4 (as this is a necessary condition for a strong RDD study). We suggest conducting the RDD alongside the follow-up of the smaller RCT, so that the evaluation continues to validate the initial sample's results with a demonstration of that the program continues to maintain its quality and effectiveness (as measured by initial impacts).

With respect to measures of quality, the standard currently is use of the CLASS. This is widely used in Washington now, and permits comparisons to other recent studies. We recommend its use to assess quality together with additional instruments. These might include the new ECERS on which there is little data. We also recommend a measure that assesses provisions for children from homes where English is not the dominant language. Also recommended are measures of how children spend their time or are engaged in the classroom, and measures that assess features of programs beyond the classroom walls, which have been largely absent from prior studies. This last is not essential, but is consistent with the design of the Seattle program and its implementation. Ideally, classroom measures would be collected in fall and spring, but some studies have conducted these assessments in spring only.

With respect to child assessments, these should include substantial numbers of demographic items for children and their families, as well as assessments of language, literacy, math, executive function, and social development. Specific assessments that have been most successfully used include: the PPVT and TVIP (for Spanish speakers), the Woodcock Johnson achievement tests (but not the vocabulary scale for which results are mixed), the HTKS, and additional brief measures of executive function (e.g., pencil tap). Teacher ratings of social and emotional development and classroom behavior can be added only if all children are in programs, or beginning with kindergarten follow-up. Pre-test should be conducted at application for the lottery, if possible; demographic data should be collected at the same time. Post-test should be conducted as late in the spring as possible. The first follow-up assessments would be conducted at kindergarten entry.

References

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

Ackerman, D. (2014) State-funded preK policies on external classroom observations: issues and status. Policy Information Report. ETS: Princeton, NJ.

Anderson, L. M., Shinn, C., Fullilove, M. T., Scrimshaw, S. C., Fielding, J. E., Normand, J., Carande-Kulis V.G., & Task Force on Community Preventive Services. (2003). The effectiveness of early childhood development programs: A systematic review. American journal of preventive medicine, 24(3), 32-46.

Arnett, J. (1989). Caregivers in day-care centers: Does training matter? *Journal of Applied Developmental Psychology*, *10*(4), 541-552.

Aughinbaugh, A. (2001). Does Head Start yield long-term benefits?. Journal of Human Resources, 36(4), 641-665.

Barnett, W. S. (2011). Effectiveness of early educational intervention. *Science*,*333*(6045), 975-978.

Barnett, W. S. (2013). Expanding Access to Quality Pre-K is Sound Public Policy. National Institute for Early Education Research: New Brunswick, NJ.

Barnett, W. S., & Masse, L. N. (2007). Comparative benefit–cost analysis of the Abecedarian program and its policy implications. *Economics of Education Review*, *26*(1), 113-125.

Barnett, W. S., Howes, C., & Jung, K. (2008). California's state preschool program: Quality and effects on children's cognitive abilities at kindergarten entry. *Final Report to the First 5 California Children and Families Commission*.

Barnett, W. S., Jung, K., Youn, M., & Frede, E. C. (2013). Abbott Preschool Program longitudinal effects study: Fifth grade follow-up. National Institute for Early Education Research: New Brunswick, NJ.

Barnett, W., Francis, J., & Youn, M. (n.d.). Rhode Island Year 2 Draft Report. New Brunswick, NJ: National Institute for Early Education Research.

Barnett, W.S., Carolan, M.E., Squires, J.H., Clarke Brown, K., & Horowitz, M. (2015). *The state of preschool 2014: State preschool yearbook*. New Brunswick, NJ: National Institute for Early Education Research.

Barnett, W.S., Jung, K., Frede, E., Hustedt, J.T., Howes, C., & Daniel-Echols, M. (2013). Effects of Eight State Prekindergarten Programs on Early Learning: A Regression-Discontinuity Analysis. National Institute for Early Education Research: New Brunswick, NJ.

Bartik, T. (2013). What does the Fitzpatrick study of Georgia's "universal" pre-K program show?. Invest In Kids. http://investinginkids.net/2013/03/05/what-does-the-fitzpatrick-study-of-georgias-universal-prek-program-show

Belfield, C. R., Nores, M., Barnett, W. S., & Schweinhart, L. (2006). The Perry Pre-School 40-year follow-up cost-benefit analysis. *Journal of Human Resources*, *41*(1), 215-246.

Boston Public Schools (2012). Early Childhood Update: Presentation to school committee. Author: Boston, MA. Retrieved from http://bpsearlychildhood.weebly.com/uploads/1/0/1/3/10131776/school_committee_4_11_12__early_childhood_presentation_3.pdf

Boston Public Schools (2014, March 10). BPS early childhood study: Program helps close achievement gaps, expands learning. Author: Boston, MA. Retrieved from http://bostonpublicschools.org/site/default.aspx?PageType=3&DomainID=4&ModuleInstanceID=14&ViewID=047E6BE3-6D87-4130-8424-D8E4E9ED6C2A&RenderLoc=0&FlexDataID=3439&PageID=1

Boston Public Schools Department of Early Childhood (2015). Boston k1ds: High quality pre-k in community-based early education programs. Author: Boston, MA. Retrieved from http://bpsearlychildhood.weebly.com/boston-k1ds.html

Bryant, D. (2010). Observational Measures of Quality in Center-Based Early Care and Education Programs, OPRE Research-to-Policy, Research-to-Practice Brief OPRE 2011-10c. Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services: Washington, DC.

Burchinal, M. R., & Cryer, D. (2004). Diversity, child care quality, and developmental outcomes. *Early Childhood Research Quarterly*, *18*(4), 401-426.

Burchinal, M., Kainz, K., & Cai, Y. (2011). How well do our measures of quality predict child outcomes? A meta-analysis and coordinated analysis of data from large-scale studies of early childhood settings. *Quality measurement in early childhood settings*, *11*.

Camilli, G., Vargas, S., Ryan, S., & Barnett, W. S. (2010). Meta-analysis of the effects of early education interventions on cognitive and social development.*The Teachers College Record*, *112*(3).

Campbell, F. A., Pungello, E. P., Burchinal, M., Kainz, K., Pan, Y., Wasik, B. H., Barbarin, O. A., Sparling, J. J., and Ramey, C. T. 2012. Adult outcomes as a function of an early childhood educational program: An Abecedarian Project follow-up. *Developmental Psychology,* 48(4), 1033-1043.

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

Cascio, E. U., & Schanzenbach, D. W. (2013). The Impacts of Expanding Access to High-Quality Preschool Education (Working Paper No. 19735). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w19735.

Castro, D. C., Espinosa, L., & Páez, M. (2011). Defining and measuring quality early childhood practices that promote dual language learners' development and learning. *Quality measurement in early childhood settings*, 257-280.

Consortium for Longitudinal Studies (Eds.) (1983). *As the twig is bent…lasting effects of preschool programs*. Hillsdale, NJ: Erlbaum.

Deutsch, M., Deutsch, C. P., Jordan, T. J., & Grallow, R. (1983). The IDS program: An experiment in early and sustained enrichment. In Consortium for Longitudinal Studies (Eds.), *As the twig is bent… Lasting effects of preschool program*s (pp. 377-410). Hillsdale, NJ: Lawrence Erlbaum.

Deutsch, M., Taleporos, E., & Victor, J. (1974). A brief synopsis of an initial enrichment program in early childhood. In S. Ryan (Ed.), *A report on longitudinal evaluations of preschool programs Volume 1: Longitudinal evaluations* (pp. 49-60). Washington, DC: Office of Child Development, US Department of Health, Education, and Welfare.

Digest of Education Statistics: 2013. (2013). Retrieved July 23, 2015, from https://nces.ed.gov/programs/digest/d13/.

Downer, J. T., López, M. L., Grimm, K. J., Hamagami, A., Pianta, R. C., & Howes, C. (2012). Observations of teacher–child interactions in classrooms serving Latinos and dual language learners: Applicability of the Classroom Assessment Scoring System in diverse settings. *Early Childhood Research Quarterly*, *27*(1), 21-32.

Duncan, G. J., & Murnane, R. J. (2014). *Restoring opportunity: The crisis of inequality and the challenge for American education* (pp. 978-1). Harvard Education Press.

Edvance Research. *Pre-K 4 SA Evaluation Report: Year 1*. (2014). Edvance Research: San Antonio, TX.

Ertem, I. O., Dogan, D. G., Gok, C. G., Kizilates, S. U., Caliskan, A., Atay, G., ... & Cicchetti, D. V. (2008). A guide for monitoring child development in low-and middle-income countries. *Pediatrics*, *121*(3), e581-e589.

Espinosa, L., & López, M. L. (2007). Assessment considerations for young English language learners across different levels of accountability. National Early Childhood Accountability Task Force and First 5: LA, CA.

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

Fitzpatrick, M.D. (2008) Starting School at Four: The Effect of Universal Pre-Kindergarten on Children's Academic Achievement. The B.E. Journal of Economic Analysis & Policy, 8(1). Available at: http://www.bepress.com/bejeap/vol8/iss1/art46

Florian, J., Schweinhart, L., Epstein, A.S (1997). Early Returns: First Year Report of the Michigan School-Readiness Program Evaluation. High/Scope Educational Research Foundation, Ypsilanti, MI.

Francis, J. (2011a) Rhode Island Prekindergarten Demonstration Project Evaluation: Year 1 Progress Report. National Institute for Early Education Research: New Brunswick, NJ.

Francis, J. (2011b) Rhode Island Prekindergarten Demonstration Project Evaluation: Year 2 Report. National Institute for Early Education Research: New Brunswick, NJ.

Frede, E. (2005). Assessment in a continuous improvement cycle: New Jersey's Abbott Preschool Program. Invited paper for the National Early Childhood Accountability Task Force with support from The Pew Charitable Trusts, the Foundation for Child Development, and the Joyce Foundation. National Institute for Early Education Research: New Brunswick, NJ.

Frede, E. C. (1998). Preschool program quality in programs for children in poverty. *Early care and education for children in poverty*, 77-98.

Frede, E. C., & Barnett, W. S. (2011). New Jersey's Abbott pre-k program: A model for the nation. In E. Zigler, W. Gilliam, & W. S. Barnett (Eds.), *The pre-k debates: current controversies and issues* (pp. 191-196). Baltimore, MD: Brookes Publishing.

Frede, E. C., Jung, K., Barnett, W. S., Lamy, C. E., & Figueras, A. (2009). The APPLES blossom: Abbott Preschool Program Longitudinal Effects Study (APPLES): Preliminary effects through second grade. National Institute for Early Education Research: New Brunswick, NJ.

Frede, E., Jung, K., Barnett, W. S., Lamy, C., & Figueras, A. (2007). The Abbott Preschool Program Longitudinal Study (APPLES) Interim Report. National Institute for Early Education Research: New Brunswick, NJ.

Freedson, M., Figueras, A., & Frede, E. (2008). Classroom Assessment of Supports for Emergent Bilingual Acquisition (CASEBA). New Brunswick, NJ: National Institute for Early Education Research, Rutgers University.

Freedson, M., Figueras-Daniel, A., Frede, E., Jung, K., & Sideris, J. (2011). The Classroom Assessment of Supports for Emergent Bilingual Acquisition: Psychometric properties and key initial findings from New Jersey's Abbott Preschool Program. In C. Howes, J. T. Downer, & R. C. Pianta (Eds.), *Dual Language Learners in the early childhood classroom* (pp. 233-258). Baltimore, MD: Paul H. Brookes

Friedman, A. H., Frede, E., Epstein, D., Sansanelli, R., Ackerman, D., & Stevenson-Boyd, J. (2009). New Jersey preschool expansion assessment research study (PEARS): Report to the state. National Institute for Early Education Research: New Brunswick, NJ

Gorey, K. M. (2001). Early childhood education: A metataly-analytic affirmation of the short-and long-term benefits of educational opportunity. *School Psychology Quarterly,* 16(1), 9.

Gormley Jr, W. T., Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal pre-K on cognitive development. *Developmental psychology,* 41(6), 872.

Gormley, W. T. (2008). The Effects of Oklahoma's Pre-K Program on Hispanic Children*. *Social Science Quarterly*, *89*(4), 916-936.

Gormley, W. T., Phillips, D., & Gayer, T. (2008). Preschool programs can boost school readiness. SCIENCE-NEW YORK THEN WASHINGTON-, 320(5884), 1723.

Graham, G. (2013, March 11). Tulsa's preschool programs seen as national model. Tulsa World: Boulder, Tulsa. Retrieved from: http://www.tulsaworld.com/news/local/tulsa-s-preschool-programs-seen-as-national-model/article_a932f79f-ec5c-5619-a89a-0ef88d271830.html

Guralnick, M. J., & Bennett, F. C. (1987). A framework for early intervention. The effectiveness of early intervention for at-risk and handicapped children, 265.

Hallam, R, Rous, B. Riley-Ayers, S & Epstein, D. (2011). Teacher survey of early education quality. New Brunswick, NJ: NIEER.

Halle, T. G., Whittaker, J. V., Zepeda, M., Rothenberg, L., Anderson, R., Daneri, P., ... & Buysse, V. (2014). The social–emotional development of dual language learners: Looking back at existing research and moving forward with purpose. *Early Childhood Research Quarterly*, *29*(4), 734-749.

Harms, T., Clifford, R. M., & Cryer, D. (2005). Early childhood environment rating scale. T*eachers College Press.*

Harms, T., Clifford, R. M., & Cryer, D. (2014). Early childhood environment rating scale. Teachers College Press.

Head Start (2015). Head Start Program Facts Fiscal Year 2014. Author: Washington DC. Retrieved from http://eclkc.ohs.acf.hhs.gov/hslc/hs/about

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A., & Yavitz, A. (2010). The rate of return to the HighScope Perry Preschool Program. *Journal of Public Economics*, 94(1), 114-128.

Hill, C., Gormley, W., & Adelstein, S. (2012). Do the short-term effects of a strong preschool program persist. Working Paper, 18, 17. Center for Research on Children in the United States: Washington, D.C.

Hustedt, J. T., Barnett, W. S., Jung, K., & Figueras, A. (2008). Impacts of New Mexico PreK on children's school readiness at kindergarten entry: Results from the second year of a growing initiative. National Institute for Early Education Research, Rutgers University.

Hustedt, J. T., Barnett, W. S., Jung, K., & Thomas, J. (2007). The effects of the Arkansas Better Chance Program on young children's school readiness. National Institute for Early Education Research.

Huston, A., Gupta, A., & Schexnayder, D. (2012). Study of Early Education in Texas The Relationship of Pre-K Attendance to 3rd Grade Test Results.

Jordan, T. J., Grallo, R., Deutsch, M., & Deutsch, C. P. (1985). Long-term effects of early enrichment: A 20-year perspective on persistence and change. *American Journal of Community Psychology, 1*3(4), 393-415.

Jung, K., Barnett, W. S., Hustedt, J. T., & Francis, J. (2013). Longitudinal effects of the Arkansas Better Chance Program: Findings from first grade through fourth grade. National Institute for Early Education Research: New Brunswick, NJ

Karoly, L. A., Kilburn, M. R., & Cannon, J. S. (2005). Proven benefits of early childhood interventions.

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

King, F. J., Cappellini, C. H., & Gravens, L. (1995). A longitudinal study of the Florida Prekindergarten Early Intervention Program, Part III. Florida State University, Educational Services Program: Tallahassee, FL.

Kuhne, Kristin. 2008. "Pre-K Effects in Texas." Unpublished Ph.D., University of Texas at Dallas: Richardson, TX.

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

Ladd, H. F., Muschkin, C. G. & Dodge, K. A. (2014). From Birth to School: Early Childhood Initiatives and Third-Grade Outcomes in North Carolina. Journal of Policy Analysis Management, 33: 162–187. doi: 10.1002/pam.21734

Lamy, C., Barnett, S. W., & Jung, K. (2005). The effects of the Michigan School Readiness program on young children's abilities at kindergarten entry: The National Institute for Early Education Research. National Institute for Early Education Research: New Brunswick, NJ.

Lipsey, M. W. (2014). Interventions for juvenile offenders: A serendipitous journey. *Criminology & Public Policy*, *13*(1), 1-14.

Lipsey, M. W., Farran, D. C., Hofer, K. G., Bilbrey, C., & Dong, N. (2011). The effects of the Tennessee voluntary pre-kindergarten program: Initial result. In annual conference of the Society for Research on Educational Effectiveness, held in Washington, DC, March (pp. 3-5). 2 other sources with same year and names but different order of names.

Lipsey, M. W., Hofer, K. G., Dong, N., Farran, D. C., & Bilbrey, C. (2013).*Evaluation of the Tennessee Voluntary Prekindergarten Program: End of Pre-K Results from the Randomized Control Design*. Research Report. Vanderbilt University, Peabody Research Institute: Nashville, TN.

Loeb, S., Bridges, M., Bassok, D., Fuller, B., & Rumberger, R. W. (2007). How much is too much? The influence of preschool centers on children's social and cognitive development. Economics of Education review, 26(1), 52-66.

Lopez, L., Arango, L., & Ferron, J. (2012, February) Positive effects of home and classroom variables on the development of bilingual language skills in preschool Latino children. Paper presented at the Society for Research in Child Development Special Topics Meeting, “Positive Development in Minority Youth”: Tampa, FL.

Ludwig, J., & Phillips, D. A. (2008). Long-term effects of Head Start on low-income children. Annals of the New York Academy of Sciences, 1136(1), 257-268.

Magnuson, K. A. & Ruhm, C. & Waldfogel, J. (2007b). Does prekindergarten improve school preparation and performance? Economics of Education Review, 26(1).

Magnuson, K. A., Ruhm, C., & Waldfogel, J. (2007a). The persistence of preschool effects: Do subsequent classroom experiences matter?. Early Childhood Research Quarterly, 22(1), 18-38.

Maloffeva, E., Daniel-Echols, M., & Xiang, Z. (2007). Findings from the Michigan School Readiness program 6 to 8 follow up study.

Maxwell, K. L., Early, D. M., Bryant, D., Kraus, S., Hume, K., & Crawford, G. (2009). Georgia study of early care and education: Findings from Georgia's Pre-K Program. Chapel Hill: The University of North Carolina, FPG Child Development Institute.

McKey, R.H., Condelli, L., Ganson, H., Barrett, B.J., McConkey, C. & Planz, M.C. (1985). The impact of Head Start on children, families, and communities. Washington, DC: Head Start Evaluation Synthesis and Utilization Project.

Mead, S. (2009). Education Reform Starts Early: Lessons from New Jersey's PreK-3rd Reform Efforts. New America Foundation: Washington, D.C.

Muennig, P., Robertson, D., Johnson, G., Campbell, F., Pungello, E. P., & Neidell, M. (2011). The effect of an early education program on adult health: the Carolina Abecedarian Project randomized controlled trial. *American Journal of Public Health*, 101(3), 512.

Muschkin, C. G., Ladd, H. F., & Dodge, K. A. (2015). Impact of North Carolina's Early Childhood Initiatives on Special Education Placements in Third Grade. *Educational Evaluation and Policy Analysis*, Month 2015, Vol. XX, No. X, pp. 1–23.

Nelson, G., Westhues, A., & MacLeod, J. (2003). A meta-analysis of longitudinal research on preschool prevention programs for children. *Prevention & Treatment*, *6*(1), 31a.

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

North Carolina Prekindergarten Program (2014). NC Pre-Kindergarten (NC Pre-K) Program Requirements and Guidance. Author: Raleigh, NC. Retrieved from http://ncchildcare.dhhs.state.nc.us/pdf_forms/NCPre-K_Program_Requirements_Guidance_2014-2015.pdf

Peisner-Feinberg, E. S., & Maris, C. L. (2005). Evaluation of the North Carolina More at Four Pre-kindergarten Program: Year 2 (July 1, 2002-June 30, 2003). Chapel Hill, NC: FPG Child Development Institute, University of North Carolina at Chapel Hill.

Peisner-Feinberg, E. S., & Schaaf, J. M. (2008). Evaluation of the North Carolina More at Four Pre-kindergarten Program: Children's Longitudinal Outcomes and Program Quality over Time (2003-2007). Chapel Hill, NC: FPG Child Development Institute, University of North Carolina at Chapel Hill.

Peisner-Feinberg, E. S., & Schaaf, J. M. (2010). Long-term effects of the North Carolina More at Four Pre-Kindergarten Program: Children's reading and math skills at third grade. Frank Porter Graham Child Development Institute: Chapel Hill, NC.

Peisner-Feinberg, E. S., LaForett, D. R., Schaaf, J. M., Hildebrandt, L. M., Sideris, J., & Pan, Y. (2014). Children's outcomes and program quality in the North Carolina Pre-Kindergarten Program:

2012–2013 Statewide evaluation. Frank Porter Graham Child Development Institute: Chapel Hill, NC.

Peisner-Feinberg, E. S., Schaaf, J. M., LaForett, D. R., Hildebrandt, L.M., & Sideris, J. (2014). Effects of Georgia's Pre-K Program on children's school readiness skills: Findings from the 2012–2013 evaluation study. Executive Summary. Frank Porter Graham Child Development Institute: Chapel Hill, NC.

Peisner-Feinberg, E. S., Schaaf, J. M., LaForett, D. R., Hildebrandt, L. M., & Sideris, J. (2014). Effects of Georgia's Pre-K Program on Children's School Readiness Skills: Findings from the 2012–2013 Evaluation Study. Frank Porter Graham Child Development Institute, University of North Carolina at Chapel Hill.

Peisner-Feinberg, E., Schaaf, J., & LaForett, D. (2013). Children's Growth and Classroom Experiences in Georgia's Pre-K Program: Findings from the 2011-2012 Evaluation Study. Frank Porter Graham Child Development Institute: Chapel Hill, NC.

Peisner-Feinberg, E.S. & Schaaf, J. M. (2007). Evaluation of the North Carolina More at Four Pre-kindergarten Program Year 6 Report (July 1, 2006–June 30, 2007). Frank Porter Graham Child Development Institute: Chapel Hill, NC.

Peisner-Feinberg, E.S. & Schaaf, J. M. (2011). Effects of the North Carolina More at Four Pre-kindergarten Program on Children's School Readiness Skills: Summary of Key Findings. Frank Porter Graham Child Development Institute: Chapel Hill, NC.

Pianta, R., Howes, C., Burchinal, M., Bryant, D., Clifford, R., Early, D., & Barbarin, O. (2005). Features of pre-kindergarten programs, classrooms, and teachers: Do they predict observed classroom quality and child-teacher interactions?. *Applied developmental science*, 9(3), 144-159.

Pianta, R., LaParo, K., & Hamre, B. (2008). The Classroom Assessment Scoring System Pre-K Manual. Baltimore: Brookes.

Pilcher, L. and Kaufman-McMurrain, M. (1996). The longitudinal study of Georgia's prekindergarten children and families. Atlanta: Department of Early Childhood Education, Georgia State University.

Puma, M., Bell, S., Cook, R., Heid, C., Broene, P., Jenkins, F., Mashburn, A, & Downer, J. (2012). Third Grade Follow-up to the Head Start Impact Study Final Report, OPRE Report #2012-45, Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. Accessed at: http://www.acf.hhs.gov/sites/default/files/opre/head_start_report.pdf

Puma, M., Bell, S., Cook, R., Heid, C., Shapiro, G., Broene, P., ... & Spier, E. (2010). Head Start Impact Study. Final Report. Administration for Children & Families.

Raden, A. (1999). Universal Prekindergarten in Georgia: A Case Study of Georgia's Lottery-funded Pre-K Program (Vol. 33). NY: *Foundation for Child Development.*

Ramey, C. T., Bryant, D. M., & Suarez, T. M. (1985). Preschool compensatory education and the modifiability of intelligence: A critical review.

Reynolds, A. J. (1995). One year of preschool intervention or two: Does it matter?. *Early Childhood Research Quarterly*, *10*(1), 1-31.

Reynolds, A. J. (1997). *The Chicago Child-Parent Centers: A longitudinal study of extended early childhood intervention*. Institute for Research on Poverty, University of Wisconsin-Madison: Madison, WI.

Reynolds, A. J. (2000). Success in early intervention: the Chicago child parent centers. University of Nebraska Press: Lincoln NE.

Reynolds, A. J., Temple, J. A., & Ou, S. R. (2010). Preschool education, educational attainment, and crime prevention: Contributions of cognitive and non-cognitive skills. *Children and Youth Services Review*, *32*(8), 1054-1063.

Reynolds, A. J., Temple, J. A., Ou, S. R., Robertson, D. L., Mersky, J. P., Topitzes, J. W., & Niles, M. D. (2007). Effects of a school-based, early childhood intervention on adult health and well-being: A 19-year follow-up of low-income families. *Archives of Pediatrics & Adolescent Medicine, 161*(8), 730-739.

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*(4), 267-303.

Reynolds, A. J., Temple, J. A., Robertson, D. L., & Mann, E. A. (2001). Long-term effects of an early childhood intervention on educational achievement and juvenile arrest: A 15-year follow-up of low-income children in public schools. *JAMA, 285*(18), 2339-2346.

Reynolds, A. J., Temple, J. A., White, B. A., Ou, S. R., & Robertson, D. L. (2011). Age 26 cost–benefit analysis of the child-parent center early education program. *Child Development*, *82*(1), 379-404.

Ritchie, S., Howes, C., Kraft-Sayre, M., & Weiser, B. (2002). Kindergarten teacher involvement scale—pre academic scale: Snapshot codebook.

Roberson, N., McClelland, A., Palaich, J. B., Rooney, K., & Workman, S. (2015) Working Draft: Long Term Persistence Effects of the Denver Preschool Program. Augenblick, Palaich and Associates: Denver, CO.

Robertson, N., McClelland, A., Palaich, B., Rooney, K., & Workman, S. (2015). Long term persistence effects of the Denver Preschool Program. In E. H. Greenberg (Chair), New Evidence on the Longer-Term Impacts of State Preschool Programs. Paper Symposium at the Association for Education Finance and Policy: Washington D.C.

Rundquist, J. (2013, March 20). Study: Preschoolers in poorest districts made 'significant gains' through 4th and 5th grades. The Star Ledger: Newark, NJ. Retrieved from http://www.nj.com/news/index.ssf/2013/03/nj_preschool_study_shows_gains.html

Sachs, J., & Weiland, C. (2010). Boston's rapid expansion of public school-based preschool: Promoting quality, lessons learned. *Young Children*, 74-77.

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.

Schweinhart, L. J., Montie, J., Xiang, Z., Barnett, W. S., Belfield, C. R., & Nores, M. (2005). Lifetime effects: The High/Scope Perry Preschool Project through age 40.

Shadish W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. *Wadsworth Cengage learning.*

Shager, H. M., Schindler, H. S., Magnuson, K. A., Duncan, G. J., Yoshikawa, H., & Hart, C. M. (2012). Can research design explain variation in Head Start research results? A meta-analysis of cognitive and achievement outcomes. *Educational Evaluation and Policy Analysis*, Vol. 35, No. 1, pp. 76–95.

Shaw, L. (2014, January 30). How Boston's preschools went from mediocre to outstanding. The Seattle Times: Seattle, WA. Retrieved from http://blogs.seattletimes.com/educationlab/2014/01/30/how-bostons-preschools-went-from-mediocre-to-outstanding/

Smith, M. W., Brady, J. P., & Anastasopoulos, L. (2008). User's Guide to the Early Language & Literacy Classroom Observation: Pre-K Tool. Paul H. Brookes Publishing Company: Baltimore, MD.

Smith, S., Davidson, S., & Weisenfeld, G. (2001). Supports for Early Literacy Assessment for early childhood programs serving preschool-age children.

Stevenson-Boyd, J. S., Brenneman, K., Frede, E., & Weber, M. (2009). Preschool rating instrument for science and mathematics. New Brunswick, NJ: National Institute for Early Education Research.

Texas Education Agency. (1995). Texas evaluation study of prekindergarten programs: Final report. Texas Education Agency. Austin, TX.

U.S. Department of Education (2014). 18 states awarded new preschool development grants to increase access to high-quality preschool programs. Author: Washington, D.C. Retrieved from http://www.ed.gov/news/press-releases/18-states-awarded-new-preschool-development-grants-increase-access-high-quality-preschool-programs

Vitiello, V. E. (2013). Research and Recommendations.

Waldfogel, J. (2004). Social mobility, life chances, and the early years. LSE STICERD Research Paper No. CASE088.

Weiland, C., & Yoshikawa, H. (2013). Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *Child Development, 84*(6), 2112-2130.

White, K., & Casto, G. (1985). An integrative review of early intervention efficacy studies with at-risk children: Implications for the handicapped. *Analysis and Intervention in Developmental Disabilities, 5*(1), 7-31.

Wong, V. C., Cook, T. D. Barnett, W. S., & Jung, K. (2008). An effectiveness-based evaluation of five state pre-kindergarten programs. *Journal of Policy Analysis and Management, 27*(1), 122-154.

WSIPP Reports. (2014). Retrieved July 23, 2015, from http://www.wsipp.wa.gov/Reports/14-01-2201

Yoshikawa, H. (2014). Investing in Our Future: The Evidence Base on Quality Preschool Education. Foundation for Child Development: New York, NY.