

**A Critical Meta-analysis of All Evaluations of State-Funded  
Preschool from 1977 to 1998:  
Implications for Policy, Service Delivery and Program Evaluation**

**Walter S. Gilliam and Edward F. Zigler  
Yale University Child Study Center**

**Abstract**

The number of state-funded preschool programs for low-income children has increased dramatically over the past few decades, and recent research has indicated that these programs vary considerably along a variety of dimensions. By 1998 only 13 of the current 33 state preschool programs (which serve children 3 to 5, provide some form of classroom-based educational service, and are primarily funded and administered at the state level) had completed a formal evaluation of the program's impact on child outcomes. This paper presents a critical meta-analytic review of these evaluations, providing measures of standardized effects for all significant impacts to facilitate comparisons across differing domains of outcome and evaluative methods. Although several methodological flaws in these studies are identified, the pattern of overall findings may offer modest support for positive impacts in improving children's developmental competence in a variety of domains, improving later school attendance and performance, and reducing subsequent grade retention. Significant impacts were mostly limited to kindergarten and first grade; however, some impacts were sustained several years beyond preschool. The results of these studies were similar to evaluations of other large-scale preschool programs for low-income children, such as Head Start. Modest outcome goals are warranted for preschool programs serving low-income children, e.g. the promotion of school readiness. Suggestions are presented for improved preschool and early intervention program evaluation.

**Suggested citation:**

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

**Address all correspondence to:**

Walter S. Gilliam, PhD; Yale University Child Study Center, 230 South Frontage Road, PO Box 207900, New Haven, Connecticut 06520-7900; Phone: 203-785-3384; Fax: 203-785-7926; E-mail: [walter.gilliam@yale.edu](mailto:walter.gilliam@yale.edu).

## A Critical Meta-analysis of All Evaluations of State-Funded Preschool from 1977 to 1998: Implications for Policy, Service Delivery and Program Evaluation

High quality preschool programs that offer a comprehensive array of child- and family-focused services have long been shown to promote school readiness and other favorable outcomes in at-risk children (Barnett, 1998; Guralnick, 1997; Karoly et al., 1998; McCall, Larsen, & Ingram, 2000; Sawhill, 1999; Zigler, 1998). Bolstered by this evidence, a wide variety of such programs has been created to meet the needs of America's most at-risk preschoolers and their families. When Head Start, the largest and most researched of these programs, was first implemented in the summer of 1965 as a federally administered program, few states were offering preschool services to their children, at-risk or otherwise. A notable exception is Wisconsin, which has had a state-funded preschool or prekindergarten program beginning as early as 1898.<sup>1</sup> One year after the federally funded Head Start program first opened its doors, the New York State Experimental Prekindergarten Program (NYSEPP)<sup>2</sup> began serving financially disadvantaged children throughout New York. Following the success of the NYSEPP in improving at-risk children's cognitive and pre-academic school readiness (Horan, Irvine, Flint, & Hick, 1980), other states began to develop their own programs, with as many as 10 state preschool programs in operation by 1980 (Ripple, Gilliam, Chanana, & Zigler, 1999). In this paper a state-funded preschool or prekindergarten program is one which serves children 3- to 5-years-old, provides some form of classroom-based educational service, and is primarily funded and administered at the state level.<sup>3</sup> (See Table 1 for the criteria we used to identify these programs.)

**Table 1**

**Definitional criteria for state-funded preschool or prekindergarten programs**

**State-funded preschool programs:**

- |     |  |
|-----|--|
| (a) | target or are accessible to children from low-income families  |
| (b) | provide at least some form of classroom-based, educational service directly to preschool-age children  |
| (c) | are mandated and administered at the state level or the District of Columbia (not state aid for low-income parents to purchase their own preschool services) |
| (d) | are primarily state-funded (not state supplementation to programs funded or administered primarily at the federal or local level)                            |
| (e) | do not serve exclusively children with disabilities  |

This definition is somewhat more restrictive than others currently being used to track state investment in early childhood care and education. Not included in this definition are the nine states that provide substantial funds to increase the capacity of federally-administered Head Start

<sup>1</sup> In 1898, Wisconsin began providing state funding for a two-year kindergarten program for children four and five years old. Funding for the four-year-old component of the two-year kindergarten was repealed in 1957. In 1985, it was restored as the Wisconsin Four-Year-Old Kindergarten Program (Schulman, Blank, & Ewen, 1999).

<sup>2</sup> The NYSEPP, implemented in 1966, should not be confused with New York's current universal prekindergarten program, implemented in 1998, which has not been formally evaluated.

<sup>3</sup> This definition is somewhat more restrictive than others currently being used to track state investment in early childhood care and education. Not included in this definition are the nine states that provide substantial funds to increase the capacity of federally-administered Head Start programs operating in their state, but otherwise do not operate their own distinctly administered and funded preschool program (Knitzer & Page, 1998; Ripple et al., 1999; Schulman et al., 1999).

programs operating in their state, but otherwise do not operate their own distinctly administered and funded preschool program (Gilliam & Ripple, in press; Knitzer & Page, 1998; Ripple et al., 1999; Schulman et al., 1999).

A boost in the number of state-funded preschool programs for low-income children occurred in the years immediately after the passage of the Individuals with Disabilities Education Act (IDEA) Amendments of 1986 (PL 99-457), which extended special education services to preschool children who meet certain disability criteria. The advent of these special education preschool programs in state departments of education offered a framework for some states to provide preschool programs for financially-disadvantaged preschoolers by expanding the criteria of eligibility to include these at-risk, albeit non-disabled, preschoolers.

Several mechanisms appear to be responsible for the dramatic increase in the number of state-funded preschool programs for at-risk children during the 1990s. In 1990 President George H. Bush and all 50 state governors endorsed what was to later become the Goals 2000: Educate America Act of 1994 (PL 103-227). The first of the six legislated educational goals for the nation is that “by the year 2000, all children will start school ready to learn,” with the first objective being that “all children will have access to high-quality and developmentally appropriate preschool programs that help prepare children for school.” Further objectives support the need to infuse these preschool programs with a comprehensive array of services, specifically mentioning health care, nutritional programs, physical education, and parent training and support programs. At about this time, some states launched massive education reform packages that included increased access to preschool, such as the case in Kentucky (Phillips, Boysen, & Schuster, 1997). Goals 2000 provided an aspirational, if not monetary, incentive for states to support preschool programs. Indeed, many new state preschool programs were born, and many other states began supplementing the budget of existing programs, such as Head Start. However, by the middle of the decade, still less than half of all low-income children in America were being provided any preschool programming (National Education Goals Panel, 1996). In addition to educational initiatives, federal and state welfare-to-work legislation has provided added incentive, as well as financial support, for increased childcare opportunities (Kaplan, 1998; United States General Accounting Office, 1999).

By fiscal year 2000, 33 states were offering their own distinct classroom-based preschool program (Gilliam & Ripple, in press; Ripple et al., 1999), and all but 9 states either administered their own preschool program or supplemented existing preschool programs operating within their state, such as Head Start (Schulman et al., 1999). Recently, Georgia (beginning 1995) and New York (beginning 1998) have made preschool open for all four-year-olds regardless of SES or other at-risk status, and are in the process of increasing funding to meet this new demand. Despite the progressive efforts of some states to move toward universal preschool, others lag far behind. State-funded preschool programs served a median 43% of the eligible children in their respective state, virtually the same percent of eligible children served nationally by Head Start (Ripple et al., 1999). However, about two-thirds of the state-funded preschool programs provide at least some of their services through contractual agreements with local Head Start grantees, a service venue second only to the public schools (Gilliam & Ripple, in press). Therefore, many children may be counted twice in these figures for state-funded preschool and Head Start. Clearly, the number of children served, especially those from financially disadvantaged families, falls woefully short of the nation’s goal of providing high quality preschool programs for all.<sup>4</sup>

---

<sup>4</sup> In fact, data indicate that children from financially-disadvantaged families are less likely to attend preschool than children from more affluent families (Hofferth, West, Henke, & Kaufman, 1994; West, Hausken, & Collins, 1993).

Nonetheless, the tide is in favor of increasing state involvement in the provision of preschool programs for at-risk children.

Programs and agencies that rely on public funds increasingly have been held accountable for demonstrating their effectiveness. State-funded preschool initiatives are no exception. In many states, formal evaluation of program implementation and impact is mandated in the state legislation authorizing the program. However, less than half of the current state-funded preschool programs have, or are currently conducting, impact evaluations of the effectiveness of their programs (Gilliam & Ripple, in press; Ripple et al., 1999). An impact evaluation is a study designed to estimate the degree to which a program improves outcomes among its participants, in this case the children who attend the preschool program. Of these state impact evaluations, only New York's (Horan, Irvine, Flint, & Hick, 1980) and aspects of the District of Columbia's (Marcon, 1999) have ever appeared in a professional journal, and few have been reviewed elsewhere. In a review of the long-term cognitive and academic impacts of both model and large-scale public preschool programs (Barnett, 1995; 1998), it was found that in many cases public programs had weaker effects than the often better implemented model programs. Further, it was concluded that Head Start was less effective than better-funded public school programs (Barnett, 1995). Since only three of the 21 evaluations of large-scale programs reviewed (District of Columbia, New York and Maryland) were of state-funded preschool programs, these findings may not be applicable to the other state-funded preschool programs not reviewed.

The proliferation of these state preschool programs raises several questions. What are the characteristics of these state-funded preschool programs, and what is the quality of their implementation? How are these state programs being evaluated? What are the effects of these state programs? How do their impacts compare to other large-scale programs, such as Head Start? In what ways might their measured impacts be associated with differences in program characteristics or evaluation methodology? The first question has been addressed elsewhere (Gilliam & Ripple, in press; Knitzer & Page, 1998; Ripple et al., 1999; Schulman et al., 1999; Smith, Fairchild, & Groginsky, 1997) and will need to be continually addressed as programs evolve. The answers to the remaining questions are largely unknown, or at least not compiled in a single, easily accessible location. The purpose of this paper, therefore, is to provide a detailed, critical review of the methods and findings of all impact evaluations of state-funded preschool programs conducted to date, with the goals of elucidating any trends in the results and generating recommendations for policy, service delivery, and program evaluation.

## ***REVIEW METHODS***

The current authors and colleagues completed a survey of the characteristics of state-funded preschool programs for low-income children, based on data collected between March 1997 and May 1998 (Ripple et al., 1999). This survey recently was updated to reflect the status of these programs during fiscal year 2000 (Gilliam & Ripple, in press), and program descriptions throughout this paper reflect these FY 2000 data. Based on information gathered as part of these surveys, state-funded preschool programs for low-income children were found to exist in 32 states, plus the District of Columbia. Of the 33 preschool programs identified, 22 programs reported having completed a formal evaluation of the effectiveness of their program. On follow-up, it was learned that 4 of the 22 misinterpreted the word "evaluation" to include such activities as site visits and statewide financial reports, 2 were only in the planning stages and had not yet begun to collect data, and 1 (Nebraska) had not yet completed its first report of findings and was unable to release any preliminary results (L. Ingram, personal communication, October 19,

1998). Furthermore, 2 states (Colorado and Minnesota) completed studies that did not yield data usable for the purposes of this paper. Specifically, Colorado<sup>5</sup> completed a study of the relative impact of different types of prekindergarten programs (Colorado Department of Education, n.d.), but it did not provide data necessary to judge the level of program effectiveness. Minnesota's evaluation was primarily focused on parent, rather than child, outcomes and included children who participated in some aspect of a complex array of services that may or may not have included an early childhood education experience (Cooke, 1992; Mueller, 1996). These two states were excluded from further review and analysis. This left a final total of 13 state programs with completed or on-going impact evaluations and at least some documentation of findings released by May 1998. Other efforts (e.g., systematic searches of research databases, internet search engines, and reference lists) to identify state programs with impact evaluations were not fruitful. Complete reports were obtained for all 13 evaluations (see Appendix), and program evaluators were contacted directly when additional information or clarification was needed. As discussed later in this paper, only 3 of these states completed process evaluations that examined the quality of implementation of these programs.

All reports were thoroughly reviewed, and specifics about the study method and findings were recorded. For the purposes of summarizing the findings, all statistical comparisons with  $p < .05$  were considered significant. In a few cases, the original evaluators did not conduct statistical tests. When necessary data (means, standard deviations, and sample sizes) were available from extant reports or interviews with evaluators, the present authors conducted Z-tests to determine statistical significance. (Cases when this occurred are indicated in the text or tables.) Because sample sizes varied greatly between evaluations (from as little as 14 matched pairs to greater than 40,000 children per group),  $p$ -level may indicate more about the evaluation's sample size than any actual observed differences between groups (Cohen, 1994). Therefore, when comparisons indicated a statistically significant difference between groups, standardized effect sizes ( $\Delta$ )<sup>6</sup> were computed using procedures described by Glass, McGaw, and Smith (1981).<sup>7</sup> Standardized effect sizes are used to measure the degree to which a program is effective at improving a particular outcome and can be compared across different programs and evaluation methods. Consistent with recent recommendations by the American Psychological Association, these standardized effect sizes were used in this paper to better understand the magnitude of program effects, evaluate the stability of impacts across programs and evaluation methodologies, and interpret results in the context of evaluations of related programs (Wilkinson & APA Task Force on Statistical Inference, 1999).

## ***A BRIEF DESCRIPTION OF THE STATE PRESCHOOL PROGRAMS EVALUATED***

A brief description of the programs reviewed in this paper is warranted prior to describing their research methods and findings. During our surveys, most state respondents reported that their

---

<sup>5</sup> The Colorado study was a pretest-posttest design with no comparison group and no follow-up, and no detailed report of findings is available (D. B. Smith, personal communication, November 4, 1998 & March, 11, 1999). Additional information is available in Fielden, Smith, Soper-Hepp, McNulty, and Randall (1994, February).

<sup>6</sup> A common convention for interpreting the magnitude of standardized effect sizes is to group them into one of four bands: trivial ( $\Delta < .20$ ), small ( $\Delta = .20$  to  $.50$ ), moderate ( $\Delta = .50$  to  $.80$ ), and large ( $\Delta = .80$  or more; Cohen, 1962, 1988). However, even effect sizes that Cohen would categorize as "trivial" can be quite meaningful when the outcomes are highly valued (McCartney & Rosenthal, 2000; Rosenthal, 1993).

<sup>7</sup> Specific formulas used to calculate effect sizes, given various types of retrievable data, are available from the first author by request.

program's primary goal was to increase school readiness among attendants, and the program was most always administered through the state's department of education. In their mission, however, these programs varied significantly in terms of their structure, accessibility, duration, classroom characteristics, comprehensiveness of services, and parent involvement efforts (Ripple et al., 1999; Gilliam & Ripple, in press). An inspection of some of these program characteristics did not reveal any remarkable differences between programs that had conducted impact evaluations and those that had not.

Most (61%) of the state prekindergarten programs in our surveys reported that local providers were required to follow well-established guidelines for early childhood care and education (e.g., Head Start Performance Standards, National Association for the Education of Young Children (NAEYC) guidelines, or some specific combination of these two guidelines). Another 16% (Arkansas, California, New Jersey, and Virginia) only required providers to satisfy state child care licensing requirements, a criteria so minimal as to have been shown to have little relationship to the provision of high quality services for young children (Young, Marsland, & Zigler, 1997). Finally, 23% of the state programs (Arizona, District of Columbia, Louisiana, Maine, Pennsylvania, West Virginia, and Wisconsin) reportedly stipulate no program guidelines specific to preschool-age children, relying only on local public school policies designed for older children. Most state programs reported recommending local providers to adhere to one of a list of suggested curricula, but none reported mandating a specific curriculum that must be followed. States most commonly endorsed NAEYC's *Developmentally Appropriate Practices* (DAP; Bredekamp & Copple, 1997) (although NAEYC does not consider DAP a curriculum) and the High/Scope approach as described in *Educating Young Children* (Hohmann & Weikart, 1995).

A few of the more easily quantified characteristics of these programs (the location of classrooms, duration in years and hours per day, degree of teacher pre-service and in-service training required, staff-child ratio, and the number of selected comprehensive services offered) are presented in Table 2. Programs reviewed here were located in a variety of settings, including the public schools. Like Head Start, state prekindergarten programs ranged in duration, with most operating on a nine-month (school-year), half-day schedule. About half of the programs, typically those primarily located in the public schools, required teachers to hold a bachelor's degree. Most other states required at least a Child Development Associate (CDA) credential.<sup>8</sup> Only Vermont did not require preschool teachers to possess any formal degree or credential. Few states required a staff-child ratio better than the 1:10 ratio that most nationally endorsed guidelines suggest as a minimum (Head Start Bureau, 1999; National Association for the Education of Young Children, 1998).

Eight comprehensive services sometimes offered by preschool programs were identified (meals with a nutritional requirement, physical health referrals, dental referrals, mental health referrals, vision and hearing testing, immunizations provided or referred, on-site family case workers, and home visits by staff). State programs reviewed here reported providing anywhere from 0 to all 8 of these services, with the median number of provided services being 4. The least provided services were on-site family caseworkers, home visits, and dental referrals. (For a more detailed description of these and other characteristics of state preschool programs, the interested reader is referred to our other papers that deal specifically with describing these programs

---

<sup>8</sup> The CDA requires teachers to possess at least: (1) a high school diploma or equivalent; (2) 480 clock hours of appropriate preschool experience; (3) 120 clock hours of specific formal early childhood education; (4) documented competency through formal observation of their teaching, satisfactory confidential evaluations from parents, and an appropriate professional resource file; and (5) passing scores on the CDA written and oral examinations (Council for Early Childhood Professional Recognition, 1996).

(Gilliam & Ripple, in press; Ripple et al., 1999), as well as reviews by Knitzer and Page, 1998; Schulman et al., 1999; and Smith et al., 1997.)

**Table 2**  
**Characteristics of State Preschool Programs with Impact Evaluations**

State	Location <sup>a</sup>	Number of Years <sup>b</sup>	Length of Day <sup>c</sup>	Min. Teacher Training <sup>d</sup>	In-Service Training?	Staff:Child Ratio <sup>e</sup>	Number of Services <sup>f</sup>	Reports Reviewed (See Appendix)
AR	V	2	F	BA	Yes	1:10	5	27
DC	PS	1	F	BA	Yes	NA	?	19-25 <sup>i</sup>
FL	V	2	F	CDA	No	1:10	2	12-17
GA	V	1	F	CDA	Yes	1:10	4	36-40
KY	V	1	H	CDA	Yes	1:10	7	51-56
LA	PS	1	F	BA	No	1:10	0	18
MD	PS	1	H	BA	No	1:10	3	26, 60
MI	V	1	H	CDA	No	1:8	5	4
NY <sup>g</sup>	*	*	*	*	*	*	*	1-3, 5-11, 57-59
SC	V	1	H	CDA	No	1:10	5	41-45
TX	V	2	H	BA	No	NA	1	47-50
VT	V	2	Local <sup>h</sup>	CC	Yes	1:10	4	46
WA	V	1	H	BA	Yes	1:6	8	28-35

*Note.* Data obtained from Gilliam & Ripple (in press) and Ripple et al. (1999), except for District of Columbia (Schulman, 1999). (a) V = Various locations; PS = Public schools only. (b) Two-year programs serve children 3 to 4 years old, and one-year programs serve only 4-year-olds. (c) Lengths of day reported are the states' minimum; F = Full school-day (about 6 or 7 hours); H = Half-day. All state programs operated for at least a full public school year (9 months). (d) BA = Bachelor's Degree; CDA = Child Development Associate Credential; CC = No formal degree or credential, but at least some college courses in child development. (e) Staff:Child ratios are only state recommendations; District of Columbia and Texas offer no state-level guidelines. (f) Each state's value represents the number of services required of each provider from the following list of 8: Meals with a nutritional requirement, physical health referrals, dental referrals, mental health referrals, vision and hearing testing, immunizations provided or referred, on-site family case workers, and home visits by staff. No data could be obtained for the District of Columbia. (g) Data for New York are not reported, since the impact evaluations reviewed in this paper were conducted on the New York State Experimental Prekindergarten Program (NYSEPP), which was very different from the current New York Universal Prekindergarten Program. Program data for the NYSEPP at the time of the impact evaluations are not available. (h) Program duration is locally determined in Vermont, with no reported state minimums. (i) Findings for the first grade and the second cohort of kindergartners in DC were obtained directly from the evaluator (R. A. Marcon, personal communication, March 7, 1999) and differ somewhat from those reported in Marcon, 1989. Data reported in the study used a much smaller sample of children than those reported in interview with the study author, due to casewise deletion of missing data when analyzing the data multivariately.

## ***METHODOLOGIES USED TO EVALUATE STATE PRESCHOOL PROGRAMS***

Of the 13 studies reviewed, 7 were conducted by third-party evaluators and 6 were conducted by the state departments of education. Third-party evaluators often were affiliated with one of the respective state's universities. Private consultants and educational research foundations also were contracted by states to evaluate their programs. Several methodological characteristics of these studies are described below and presented in Table 3.

### ***Number of Cohorts, Length of Follow-up and Number of Subjects***

Most states evaluated multiple cohorts of children. Kentucky had evaluated the most cohorts, with plans to evaluate additional cohorts. Most states followed participants prospectively beginning in their prekindergarten year. Some, however, recruited children once they enrolled in kindergarten, when a comparison group could be located most easily. The median length of follow-up was third grade. Maryland followed children until their tenth grade, the farthest of the state evaluations. The number of subjects varied significantly by state, since some evaluations relied on individual assessment of representative samples, whereas others relied mostly on school-collected data that existed for all students. To recruit samples, evaluators selected school districts or buildings to represent various regions of the state and then randomly selected subjects at the classroom level. Most study attrition rates ranged from less than 10% to

about 25% per year, rather similar to the 20% rate typical for evaluations of programs serving at-risk families (Gomby, 1999).

**Table 3**  
**Methodological Characteristics of State Preschool Evaluations**

State	No. of Cohorts	Pretest?	Grade Levels Assessed	Comparison Group?	Type of Comparison Group <sup>a</sup>	Sample Size <sup>b</sup>
AR	1	Yes	PreK	No		P = 175 (108 Center-based; 67 Home-based)
DC	4	No	PreK-5	Yes	M <sup>f</sup>	Cohort 1: 22 pairs; Cohort 2: 112 pairs; Cohort 3: 234 pairs; Cohort 4: 202 pairs
FL	3	Yes	PreK-4	Yes	E <sub>no</sub>	Cohorts 1-3: P ≈ 500-700, C ≈ 400-700
GA	2	No	K-1	Yes	M <sup>f</sup>	Cohort 1: P = 111, C = 111; Cohort 2: P = 267, C = 267
KY	6	Yes	PreK-4	Yes	Cohorts 1, 2, 6 = RC Cohorts 3-5 = E <sup>c</sup>	Cohorts 1-6: P ≈ 120-320, C ≈ 30-200
LA	2	Yes	PreK	No		Nearly all enrolled children (~ 1500 each cohort)
MD	1	No	K; 3; 5; 8; 9 & 10	Yes	RC	P = 416, C = 476
MI	1	No	K	Yes	E <sub>no</sub> <sup>d, f</sup>	P = 351, C = 279
NY	5	Yes	PreK-6	Yes	WL <sup>e</sup>	Cohorts 1-5: P ≈ 1000-1900, C ≈ 57-1700 <sup>h</sup>
SC	3 <sup>g</sup>	No	1-3	Yes	Cohort 1 = E Cohort 3 = RC	Cohort 1: P ≈ 600-721, C ≈ 2500-4700; Cohort 3: P ≈ 3000, C ≈ 4000
TX	1	No	PreK-3	Yes	E	P ≈ 1499-46,379, C ≈ 396-43,589 <sup>h</sup>
VT	1	No	K-2	No		P = 280
WA	3	No	PreK-7	Yes	M	Cohort 1: 250 pairs; Cohort 2: 156 pairs; Cohort 3: 946 pairs

*Note.* (a) M = Program eligible non-attendees matched to participants on some characteristics; E = Program eligible non-attendees; E<sub>no</sub> = Program eligible non-attendees who have attended no other form of preschool; WL = Program eligible children who were placed on a wait list; RC = Random classmates that may or may not be comparable to participants. (b) P = Prekindergarten Participants; C = Comparisons. (c) In cohorts 3 through 5, Kentucky recruited a comparison group of program eligible non-attendees. Due to difficulties with attrition, however, the eligible non-attendees were replaced in each cohort as needed with random classmates beginning in kindergarten, gradually transforming the comparison group from E to RC in follow-up years. (d) Participants must have attended the program 100 of 151 program days. (e) Two types of “wait listed” children were used depending on the variable and cohort: 1) children who were eligible for the program and applied, but were placed on a wait list due to a lack of space; 2) children who were eligible for the program, but were from other school districts that yet had not implemented the program. (f) Comparison group was recruited in kindergarten. (g) Only results for the first and third cohorts of the South Carolina evaluation could be located. (h) Exact number of comparison children depended on the specific outcome being measured.

### Comparison Groups

All but three of the state evaluations used some form of comparison group, against which program impacts were estimated. All three of these studies without comparison groups are severely limited methodologically and will be described in their own section later in this paper. No state evaluation randomly assigned children to program and control groups, and therefore all resorted to some less rigorous comparison group.

Five different types of comparison groups were used: waitlisted comparisons, matched program-eligible non-attendees, non-matched program-eligible non-attendees, program-eligible non-attendees who did not attend any other preschool program, and random classmates that may or may not have been eligible for the program. Some evaluations used multiple types of comparison groups. Of course the type of comparison group used has a profound effect on the interpretation and validity of the findings.

Arguably, of the comparison groups used, the *waitlist comparison* provided the best test of the program, since comparison children and their families were both eligible for services and motivated to apply for the program. Unfortunately, the only program to employ this comparison was New York's, which was by far the most outdated state evaluation (University of the State of New York & New York State Education Department, 1977). Use of the three other comparison groups of program-eligible non-attendees may have introduced a motivational bias to the results, since families of comparison children were not motivated to seek placement in the program

being evaluated. The *matched program-eligible non-attendee comparison* attempts to methodologically control some of this bias by matching the two groups on related variables (e.g., gender; ethnicity; parent education and occupational level; and some type of proxy for family income, such as eligibility for free or reduced lunch at school). The *program-eligible non-attendee comparison* is less rigorous, since comparison children, though eligible for the program, may have significantly differed from participants in many ways. The comparison group that arguably provides the least stringent test of the program is the one that used *program-eligible non-attendees who did not attend any other preschool program*, since the families of comparison children were not motivated to seek preschool programming for their child. Furthermore, in Michigan the participants were required to have attended the program at least two thirds of the total number of program days (100 of 150 days). Stipulating that the participant children must have attended the program at least minimally and that the comparison children could not have participated in a similar program may bias studies toward finding positive results. However, one could also argue that it is unfair to test the effectiveness of a program by using participants who have not participated in at least some minimal way or by comparing outcomes to children who may have attended similar programs (Gilliam, Ripple, Zigler, & Leiter, 2000).

Three evaluations used *random elementary school classmates* as comparisons. Despite this method of selecting comparisons, the two groups in Maryland were comparable in ethnicity, age, gender, family composition, and father's educational and occupational level, but differences in maternal educational and occupational level slightly favored preschool participants. In Kentucky and South Carolina, random classmates comparisons were used only for certain cohorts or at certain grade levels. Additionally, in 3 of Kentucky's 6 cohorts a program-eligible non-attendee comparison group was initially recruited, but then changed to random classmates over the course of yearly follow-ups as a result of serious attrition rates in the comparison group. This practice, of course, makes it difficult to determine the nature of the comparison group in the follow-up years and virtually precludes a longitudinal treatment of the data.

### ***Domains of Outcomes Assessed and Instruments Used***

Outcomes were categorized in one of 11 domains, and most states tracked outcomes in more than one domain (see Table 4). These outcome domains include developmental competence, children's self-perceived competence or self-efficacy, behavior problems, physical health, school attendance, grades, academic achievement test results, grade retention, special education referral and placement, parent involvement during elementary school years, and drop out during high school. Developmental competence included measures of social-emotional development, self-help skills, motor skills, language skills, cognitive development, and academic and literacy skills. Usually, developmental tests provided an estimate of children's functioning in one or more of these areas of development. After developmental competence, special education referral/placement and test scores on group administered school achievement tests (e.g., the California Achievement Test and the Iowa Test of Basic Skills) were the most commonly measured outcome domains. These domains are described in the section below that addresses evaluation findings. A large number of tests and procedures were used in these 13 evaluations. Several of them are well-known, psychometrically valid instruments. In many cases, however, relatively unknown tests were used with little data regarding their reliability and validity.

**Table 4**  
**Outcomes Assessed and Instruments Used in Evaluations of State Preschool Programs**

State	Outcomes Assessed										
	Developmental Competence	Perceived Competence	Behavior Problems	Child Health	Attendance	Grades	Achievement Tests	Drop Out	Retention	Special Ed	Parent Involvement
AR	23										
DC	40		40			33	7				
FL	22, 42, 4, 16		34		34	33	5, 7, 28, 37, 15		34	34	22
GA	17, 18				22		24		22	22	22
KY	2, 3, 25, 35, 36	20	36		22					22	
LA	14										
MD	27				34		5, 26	34	34	34	
MI	10, 21										
NY	11, 13, 29, 41				22		32		22	22	
SC	12						1		34	34	
TX							39		22	22	22
VT	22					22			22	22	
WA	8, 16, 29		9	22		22				22	22

**Instruments Used**

1	<i>Basic Skills Assessment Program</i>	16	<i>Developmental Indicators for the Assessment of Learning-Revised</i>	30	<i>Preschool Developmental Inventory</i>
2	<i>Battelle Developmental Inventory</i>			31	<i>Preschool Language Scale</i>
3	<i>Book Handling Test</i>	17	<i>Developmental Profiles-II</i>	32	<i>Pupil Evaluation Program Test</i>
4	<i>Brigance Diagnostic Inventory</i>	18	<i>Developmental Rating Scale</i>	33	Report cards
5	<i>California Achievement Test</i>	19	<i>Early Screening Inventory</i>	34	School records
6	<i>California Test of Basic Skills</i>	20	<i>Harter Scale of Perceived Self-Competence</i>	35	<i>Sentence Repetition Test</i>
7	<i>California Preschool Social Competency Scale</i>	21	<i>High/Scope Child Observation Record</i>	36	<i>Social Skills Rating System [Teacher &amp; Parent Forms]</i>
8	<i>Child Adaptive/Student Behavior Inventory</i>	22	<i>Homemade surveys, interviews, or tests</i>		
9	<i>Child Behavior Inventory</i>	23	<i>Individual Developmental Early Assessment</i>	37	<i>Stanford Achievement Test</i>
10	<i>Child Development Rating</i>	24	<i>Iowa Test of Basic Skills</i>	38	<i>Test of Early Language Development</i>
11	<i>Cognitive Abilities Test</i>	25	<i>Letter Recognition Test</i>	39	<i>Texas Assessment of Academic Skills</i>
12	<i>Cognitive Skills Assessment Battery</i>	26	<i>Maryland Functional Test</i>	40	<i>Vineland Adaptive Behavior Scales [Classroom Version] (Age Normed)</i>
13	<i>Cooperative Preschool Inventory</i>	27	<i>Maryland Systematic Teacher Observation Instrument</i>	41	<i>Walker Readiness Test for Disadvantaged Children</i>
14	<i>Creative Curriculum Checklist</i>				
15	<i>Comprehensive Achievement Program</i>	28	<i>Metropolitan Achievement Test</i>		
		29	<i>Peabody Picture Vocabulary Test-Revised</i>	42	<i>Yellow Brick Road</i>

**FINDINGS FROM STATE PRESCHOOL EVALUATIONS**

Tables 5 and 6 indicate the standard effect size for all statistically significant ( $p < .05$ ) analyses, organized by domain of outcome, grade level and state. Since only statistically significant findings were considered to be reliable enough to report, only effect sizes representing statistically significant analyses are presented numerically. Null, as well as negative, findings are also noted, but with no accompanying effect size. In many cases, effect sizes from more than one cohort are represented in a single cell, starting with the earliest cohort. Only findings from the 10 state evaluations that used a comparison group of some kind are presented. The three state evaluations that did not use a comparison group are described later in this paper. These effect sizes, presented in the tables and described below, should be interpreted cautiously, bearing in mind the methodological limitations described throughout this paper. Consideration of these effect sizes will lead the discussion of findings, since they represent the best estimate of the magnitude of program impacts.

**Developmental Competence**

Twelve of the 13 states (all except Texas) gathered at least some data on children’s developmental competence. Evaluation results are reported both for tests that address specific subdomains of developmental competence (social, self-help, motor, language, cognitive, and academic or literacy skills) and tests that provide an overall developmental score that combines two or more specific subdomains of development. Impacts on overall developmental scores are presented in Table 5, and impacts for specific subdomains are presented in Table 6.

**Table 5**  
**Standardized Effect Sizes of Statistically Positive Impacts of State Preschool Programs Through Grade 4**

	End PreK	K	1	2	3	4
<b>OVERALL DEVELOPMENT</b>						
DC		.29/ns	.93/ns			
FL	1.65/1.31	.44				
GA			ns			
KY	.22/.32/.32	.33/ns/ns	ns/ns	ns		
MD		.53				
MI		.27&.51 <sup>h</sup>				
NY	.44&.61/.48&.44 <sup>h</sup>	.42&ns <sup>h</sup>			ns	
SC			.15			
WA		NR	NR	NR		
<b>PERCEIVED COMPETENCE</b>						
KY	.18	ns	ns			
<b>BEHAVIOR PROBLEMS</b>						
FL						.33
KY	ns/.32/ns <sup>i</sup>	ns/ns/ns/ns/ns	ns/ns/ns/-	-/ns/ns/ns	-/ns	ns
WA			ns <sup>a</sup>	ns <sup>a</sup>	ns <sup>a</sup>	
<b>CHILD HEALTH</b>						
WA			ns <sup>a</sup>	ns <sup>a</sup>	ns <sup>a</sup>	
<b>ATTENDANCE</b>						
FL		ns/.15	ns/.16	ns/ns	ns/.13	
GA		.17	ns/.18			
KY		ns <sup>a</sup> /ns <sup>a</sup>	ns <sup>a</sup> /ns <sup>a</sup>	ns <sup>a</sup> /ns <sup>a</sup>	ns <sup>a</sup> /ns <sup>a</sup>	ns <sup>a</sup> /ns <sup>a</sup>
NY		.09	.12	.13	.13	.13
<b>GRADES</b>						
DC (Reading)		.30/ns	ns/ns		ns	ns
DC (Math)		.37/ns	ns/ns		ns	ns
FL (Reading)		ns	ns	ns	-	ns
FL (Math)		ns	ns	ns	-	ns
WA (Reading)			ns <sup>a</sup>	ns <sup>a</sup>	ns <sup>a</sup>	
WA (Math)			.38 <sup>a</sup>	ns <sup>a</sup>	ns <sup>a</sup>	
<b>ACHIEVEMENT TESTS</b>						
DC (Reading)					ns	
DC (Math)					ns	
FL (Reading)		.23	-/ns	ns/ns	ns/ns	-
FL (Math)		.25	ns/ns	ns/ns	ns/ns	-
GA (Reading)			.24			
GA (Math)			ns			
MD (Reading)					.39	
MD (Math)					.50	
NY (Reading)					ns	
NY (Math)					.16	
SC (Reading)			.07 <sup>b</sup> /ns <sup>c</sup>	ns <sup>c</sup>	ns <sup>c</sup>	
SC (Math)			.07 <sup>b</sup> /ns <sup>c</sup>	ns <sup>c</sup>	ns <sup>c</sup>	
TX (Reading)					.08	
TX (Math)					.09	
<b>RETENTION</b>						
FL		.14	ns	-	ns/.20 <sup>d</sup>	ns
GA		.38	ns/ns			
MD <sup>c</sup>					.52	
NY		ns	.31	ns	ns	ns
SC			1.81	ns	- <sup>a</sup>	
TX				.20		
<b>SPECIAL ED REFERRAL</b>						
GA		ns	ns/ns			
TX				.12		
WA		ns <sup>a</sup>				
<b>SPECIAL ED PLACEMENT</b>						
FL		ns/ns	ns/ns	ns/ns	ns/ns	ns
NY		ns	ns	ns	ns	ns
SC			.27			
TX				.12		
WA		ns <sup>a</sup>				

*Continued on next page.*

**Table 5 (continued)**

	End PreK	K	1	2	3	4
<b>PARENT INVOLVEMENT</b>						
GA		ns	ns/ns			
TX			.15			
WA <sup>g</sup>			ns <sup>a</sup>	ns <sup>a</sup>	ns <sup>a</sup>	

*Note.* A number indicates the standard effect size of a statistically significant positive impact ( $p < .05$ ) in favor of prekindergarten participants; “ns” indicates no significant difference between prekindergarten participants and comparison children; “-” indicates a statistically significant difference ( $p < .05$ ) in favor of comparison children; “NR” indicates that not enough data could be obtained to determine either significance or effect size. Standard effect sizes were computed by the authors of this paper using formulas presented by Glass et al. (1981).

(a) Significance level was computed by this author using a Z-test of data presented in each state’s reports.  
 (b) Comparison group consisted of program eligible non-attendees.  
 (c) Comparison group consisted of random classmates who may or may not have been eligible for the program.  
 (d) The positive finding was obtained in a second cohort of students who were analyzed for a grade retention that occurred anywhere from grade K to 3.  
 (e) Maryland data were analyzed cumulatively (e.g., at least one grade retention by third grade, by fifth grade, etc.).  
 (f) Data were analyzed cumulatively (e.g., placement in special education by fifth grade).  
 (g) Parent involvement was assessed by both parent and teacher report. In both cases results indicated no significant differences at all grade levels.  
 (h) Multiple instruments were used to measure the same construct for a single cohort.  
 (i) Effects in all 3 cohorts were based on parent reported behavior problems pretest to posttest.

**Table 6**  
**Standardized Effect Sizes of Statistically Positive Impacts in Various Subdomains of Developmental Competence Through Grade 5**

	End PreK	K	1	2	3	4	5
<b>SOCIAL</b>							
DC		ns/ns	ns/ns				ns
GA		.30	ns/ns				
KY	.22/.32/.32	ns/ns/ns/ns/ns	ns/ns/ns/-	-/ns/ns	-/ns	ns	
MI		.45					
<b>SELF-HELP</b>							
DC		ns/ns	1.23/ns				ns
GA		.41					
KY	.22/ns	.33/ns/ns	ns/ns	ns			
<b>MOTOR</b>							
DC (Overall)		ns/ns					
GA (Overall)		.45					
KY (Gross)	ns/ns	ns/ns/ns	ns/ns	ns			
KY (Fine)	.22/.32	ns/ns/ns	ns/ns	ns			
<b>LANGUAGE</b>							
DC (Overall)		.30/ns	.93/ns				ns
GA (Overall)		.39					
KY (Receptive)	ns/ns	ns/ns/ns	.40/ns	ns			
KY (Expressive)	.22/ns	ns/ns/ns	ns/ns	ns			
MI (Overall)		.45					
NY (Receptive)	ns/.42	ns			ns		
WA (Receptive)	.40/.36/.40						
<b>COGNITIVE</b>							
KY	.22/ns	ns/ns/ns	ns/ns	ns			
NY					ns		
SC			.15/-				
<b>ACADEMIC/LITERACY</b>							
GA		.42	.21				
KY (Academic)		.39/ns/ns/ns	.44/ns/ns	-/ns	ns		
KY (Literacy)	ns/ns	ns/ns	.67/ns				

*Note.* A number indicates the standard effect size of a statistically significant positive impact ( $p < .05$ ) in favor of prekindergarten participants; “ns” indicates no significant difference between prekindergarten participants and comparison children; “-” indicates a statistically significant difference ( $p < .05$ ) in favor of comparison children. Standard effect sizes were computed by the authors of this paper using the formulas presented by Glass et al. (1981).

In general, effects in overall developmental competence were both sizable and robust. In all instances, significant positive impacts were reported by the end of preschool, with consistently non-trivial effect sizes measured by Cohen’s convention.<sup>6</sup> Additionally, in all state evaluations non-trivial positive effects were sustained to kindergarten for at least one cohort.

Significant effects were inconsistent at first grade and nonexistent beyond that point, though only two states meaningfully evaluated outcomes in this domain beyond first grade.

Data presented in Table 6 suggest that the positive impacts in overall developmental competence may be attributed to short-term impacts across a variety of developmental subdomains. By the end of preschool and into kindergarten, significant impacts were found in almost every subdomain assessed. By first grade, significant effects were rare, but most frequently occurring in the language and academic/literacy domains. Significant positive effects were not noted in any developmental subdomain beyond first grade, with some negative effects in Kentucky when preschool participants were compared to random classmates.

### ***Child Perceived Self-Competence***

Kentucky was the only state to collect data on how preschool participants perceived themselves after participating in the program. At the end of the prekindergarten year, participants perceived themselves to be significantly more competent in the cognitive domain, relative to program eligible non-attendees. No significant differences, however, were observed in kindergarten or first grade, when participants were compared to random classmates.

### ***Behavior Problems***

Four states evaluated program impacts on children's behavior problems. Neither Kentucky nor Washington reported significant impacts beyond preschool using teacher rating scales, and DC (not indicated in the table) found no significant impacts at fifth grade. Florida, however, did report a significant impact as late as fourth grade. In contrast to other states, Florida relied on actual reported incidents of corporal punishment, in-school and out-of-school suspensions, and expulsions, as reported in school records. Combining data over four Florida counties, eligible non-attendees with no preschool experience (32%) were significantly more likely than participants (11%) to have been disciplined during the school year.

### ***Child Health***

Washington was the only state to collect data on this variable. A 12-item questionnaire was used to ask parents about the health of their child. Z-tests computed by the present authors using reported data indicated non-significant differences between preschool participants and matched comparisons at first, second, and third grades ( $p = .83$ ;  $p = 1.00$ ;  $p = .99$ , respectively).

### ***Attendance***

With the exception of Kentucky, all states evaluating this outcome found significant impacts in one evaluation cohort. Furthermore, these effects persisted well beyond school entry. In addition to the effects shown in Table 5, New York found statistically significant impacts at fifth and sixth grades ( $\Delta s = .13$ ), and Maryland reported a sizable positive impact at tenth grade ( $\Delta = .47$ ). The Kentucky analyses, in contrast, compared participants to random classmates who were known not to be comparable to the treatment group. The Kentucky evaluators interpreted this lack of significant difference as being indicative of a positive effect, since participants performed similarly to their less at-risk classmates.

### ***Grades***

Although report card grades for a variety of subjects were reported by different evaluations, only grades in reading and math are reviewed here. At the younger grade levels,

grades for subjects such as “verbal skills” were considered to be relatively synonymous to reading. Statistically significant impacts in DC were only found in kindergarten, and then only in the first of two cohorts. The second DC kindergarten cohort also found non-trivial effects in both reading ( $\Delta = .23$ ) and math ( $\Delta = .26$ ), but with  $N = 47$  pairs analyses did not have enough statistical power to reach significance. Washington found a significant impact only for math and only at the first grade. Statistically significant positive impacts were not reported for Florida from kindergarten through fourth grade. Interestingly, both DC and Washington found effects using a matched comparison group, whereas Florida, which did not find effects, relied on a comparison group that was not matched to participants.

### ***School-Administered Academic Achievement Tests***

Similar to report card grades, academic achievement test scores were available for a variety of domains and sub-domains. Again, for the sake of parsimonious presentation, only overall reading and overall mathematics test score findings are presented here. With the exception of DC’s, all seven evaluations addressing this outcome reported statistically significant impacts on academic achievement tests occurring at one or more grade levels. Standardized effect sizes in some cases were relatively low, however. The relative consistency of statistically significant findings for this outcome may have been due, at least in part, to the large number of subjects involved in many of these analyses. For example, statistically significant findings were reported in both South Carolina’s first grade and Texas’ third grade, with effect sizes raging only from .07 to .09. In both states, however, sample sizes were several thousand large. Conversely, DC reported no statistically significant impacts in third grade scores, despite effect sizes about three times as large ( $\Delta = .23$ ). DC’s sample size, however, was only 29 matched pairs, resulting in weak statistical power. The non-significant findings in South Carolina all occurred when participants were compared to random classmates.

Maryland and New York evaluated impacts in this domain beyond fourth grade, finding statistically significant positive impacts at every grade level assessed by either state. Maryland found significant impacts in both reading and math in fifth ( $\Delta = .40$  and  $.46$ , respectively), eighth ( $\Delta = .34$  and  $.49$ , respectively), ninth ( $\Delta = .41$  and  $.29$ , respectively), and tenth ( $\Delta = .30$  for math only) grades. New York found significant impacts in both reading and math in sixth grade ( $\Delta = .13$  and  $.12$ , respectively).

### ***School Drop Out***

The only state evaluation to collect data long enough after prekindergarten attendance to be able to address this important issue was Maryland. By tenth grade, 8.2% of the prekindergarten participants already had dropped out of school, compared to 11.3% of comparison children. Although the difference was not statistically significant, the standardized effect size approached a degree of meaningfulness ( $\Delta = .18$ ), as defined by Cohen. Since drop out was not assessed beyond tenth grade, it is unknown whether the effect would have been significant by the end of what should have been each student’s twelfth grade.

### ***Retention Rates***

In many ways positive impacts in this outcome may be one of the most robust findings for state programs, since every state that evaluated this outcome found a statistically significant impact at one or more grade levels, and effect sizes were almost always non-trivial by Cohen’s conventions. Grade level retention data were reported and analyzed in the evaluations either

cumulatively or non-cumulatively. *Non-cumulative* reporting of grade retention means that the data reflect the percentage of children retained *at an actual grade level*. For instance, a 3% non-cumulative retention rate at third grade would mean that 3% of the children enrolled in third grade for a given school year were retained. *Cumulative* reporting, however, means that the data reflect the percentage of children that were retained *by a given grade level*. For instance, a 9% cumulative retention rate at third grade would mean that 9% of the children who should have been enrolled in third grade by virtue of their chronological age had been retained at some point prior to third grade.

All evaluations except Maryland and the second cohort of Florida were analyzed non-cumulatively. When data were analyzed *non-cumulatively*, positive impacts were always found in each state's evaluation at one grade level (either kindergarten, first grade, second grade, etc.), with each state's mean  $\Delta$  across all grade levels ranging from .02 to .48 (median state mean  $\Delta = .20$ ). Additionally, when Maryland and Florida analyzed the data *cumulatively* at the third grade, both states found significant positive impacts of non-trivial size ( $\Delta s = .52$  and  $.20$ , respectively). Therefore, the mixed findings at specific grade levels when analyzed non-cumulatively may say more about each state's policy regarding when to retain children and less about the apparent robustness of the findings. Maryland also evaluated cumulative retention rates beyond fourth grade, finding significant impacts at fifth ( $\Delta = .58$ ), eighth ( $\Delta = .54$ ) and tenth ( $\Delta = .52$ ) grades. The absolute differences in cumulative retention rates for Maryland participants and comparisons were 26% versus 45% at third grade, 28% versus 50% at fifth grade, 34% versus 55% at eighth grade, and 44% versus 64% at tenth grade.

### ***Special Education Referral and Placement Rates***

Overall, few significant differences were reported for this outcome. Although most evaluations analyzed this outcome non-cumulatively, Maryland was the only state to examine this variable cumulatively. By fifth grade, only 13% of Maryland preschool participants reportedly had been placed in special education services at some point in their schooling, as compared to 24% of comparisons ( $\Delta = .42$ ;  $p < .01$ ). As was the case with grade retention data, it appears that cumulative analyses of this variable may be more likely to yield significant positive impacts than dividing the effects over successive grade levels in non-cumulative analyses.

### ***Parent Involvement***

Only three states collected data on parental involvement in their child's subsequent elementary education. Only Texas reported a statistically significant positive impact ( $\Delta = .15$ ). Georgia also found an effect of .15 in one cohort of first graders, but the sample size did not allow enough power for the results to reach statistical significance.

## ***METHODS AND FINDINGS FROM EVALUATIONS WITHOUT COMPARISON GROUPS***

As previously presented, three state evaluations (Arkansas, Louisiana, and Vermont) did not use a comparison group. All three of these evaluations have serious methodological limitations that far exceed those of the other state evaluations reported earlier. *Arkansas* reported a statistically significant improvement for participants in overall developmental competence from pretest to posttest. However, the results seem to be gravely biased, since the children's preschool teachers administered both pretests and posttests (apparently recording the results of each item on the

same score sheet) and used a form similar to the test protocol to plan individualized lessons for the students. These practices may have introduced a high level of administrator bias and “teaching to the test.” *Louisiana* used a test without norms in a single group, pretest-posttest design (L. Urbatsch, personal communication, July 28, 1998). Although results indicated significant improvements across all domains in both cohorts, these results could most likely be attributed to simple maturation, rendering the findings relatively moot as an indication of program effectiveness. The *Vermont* study used a simple single-group posttest-only design that asked elementary teachers to rate former preschool participants in relationship to their peers. The results of Vermont’s study simply do not provide the kind of data necessary to estimate program impacts.

## ***IMPLICATIONS OF THE FINDINGS***

These evaluation efforts ranged from simple to highly complicated. The 10 evaluations for which effect sizes were reported in this paper provide findings that may be used to build modest evidence for program effectiveness in certain outcome domains. In contrast, the three evaluations described in the preceding section were so limited or flawed that little regarding program impacts can be learned. These methodologically fatal flaws include practices that build in a strong evaluator bias, as well as basing results solely on single-group pretest-posttest analyses without attempting to control for maturation and other factors. Several implications for understanding the impacts of these programs and methods for evaluating them follow. Recommendations for program evaluation in the area of early childhood education are discussed and summarized in Table 7.

---

**Table 7**  
**Recommendations for the Evaluation of Preschool Program Effects**

1. Choose evaluation outcomes that are related to the goals of the program as it is implemented and are realistic given the evaluative findings for other similar programs.
  2. When random assignment to preschool and control group is not possible, evaluators should attempt to use the most comparable contrast groups possible and control for baseline differences either methodologically or statistically.
  3. Only use psychometrically reliable and valid tests.
  4. Avoid blatant misuses of null hypothesis significant testing.
  5. Present results of statistical analyses both in terms of inferential test results and standardized effect sizes, in order to address both the reliability of the findings and the magnitude of the effect.
  6. Reduce the likelihood of spurious positive findings, by using multivariate analyses or one of several Bonferroni-based correction procedures.
  7. Obtain samples of sufficient size to detect meaningful effects.
  8. Event-based outcomes (e.g., grade retention, special education placement) often should be analyzed cumulatively over time or by accounting for the amount of time before the event through survival analyses.
  9. Minimize attrition, especially methodologically produced attrition that is biased toward either preschool or contrast groups, and test for the effects of selective attrition.
  10. Use process evaluations to determine whether the program is ready for an impact evaluation and to place the results of the impact evaluation within the context of program quality.
- 

## ***What are the Impacts of These Programs, and What Outcomes Should be Measured?***

Several positive impacts were reported in a variety of outcome domains. Effects were found in several areas of developmental competence and achievement tests, extending into

kindergarten and sometimes beyond. Reduced grade retention appeared to be a rather robust impact, with cumulative effects that may last well beyond elementary and middle school. Given that grade retention is often associated with a variety of poor academic outcomes for children (Heubert & Hauser, 1999), this is a particularly noteworthy finding.

Surprisingly, positive impacts were seldom observed in the domains of special education referral and placement, parent involvement, and social development and behavioral problems. The potential for preschool programs to serve as a preventive for later delinquency has attracted attention (Yoshikawa, 1995; Zigler, Taussig, & Black, 1992), and the general lack of positive impacts in this area is interesting. One possible explanation is that behavior-rating scales may not be an adequate method for documenting this outcome. Indeed, two of the studies largely responsible for current interest in this area (Lally, Mangione, & Honig, 1988; Schweinhart, Barnes, & Weikart, 1993) relied on actual reports of subsequent delinquent behaviors, rather than more subjective teacher or caregiver ratings. Florida, the only state to find significant impacts beyond the program year, was the only state evaluation that used actual records of punishment for behavioral problems, as opposed to teacher or caregiver ratings of child behavior.

### ***How Long Do Impacts Last, and How Far Should We Reasonably Expect to Find Effects?***

Undoubtedly encouraged by the positive impacts of model preschool programs that have been known to extend well into participants' school years and beyond (Barnett, 1995; 1998), several state preschool evaluators have also attempted to track impacts well beyond the intervention. Indeed, Maryland and New York have been successful at finding these impacts at the middle and high school grades. With the exception of these two states, however, most statistically significant impacts generally were sustained only as far as kindergarten or first grade. In very few instances were statistically significant impacts not found at early grades, only to emerge later. The pattern of findings presented here more closely resembles the modest effects of other large-scale programs, like Head Start (McKey et al., 1985), rather than the more impressive impacts of smaller-scale model programs.

These findings question the utility of holding preschool programs accountable for sustaining impacts beyond kindergarten or first grade. Since the primary stated goal of most all of these state preschool programs is the promotion of school readiness (Ripple et al., 1999), evaluations of preschool program should arguably focus chiefly on impacts at the time of school entry (Zigler, 1998). Similarly, according to its most recent reauthorization, Head Start's primary goal is the promotion of school readiness (Coats Human Services Reauthorization Act of 1998). Judged by the criteria of school readiness, the evaluations reviewed here generally provide relatively consistent evidence of effectiveness, as measured by children's improved developmental competence, school attendance and school test scores and reduced grade retention. However, these findings are limited by the often inadequate and inconsistent research methods, measurements, and analyses used, as further described below.

### ***How Should State-Funded Preschool Programs be Evaluated?***

***The importance of a well-defined counterfactual.*** One of the most striking differences among these evaluations is the variety of comparison groups used. No state evaluation used a randomly selected control group, and of those that used some other form of contrast group, only a few matched the groups in terms of their at-risk status or at least pretested both groups to statistically control for any baseline differences. As indicated earlier, evaluations that used

matched comparison groups sometimes found significant impacts, when evaluations that used less stringent and more convenient comparisons did not. Clearly, there are circumstances where random assignment may not be feasible. However, evaluators must endeavor to use a comparison group that is as similar as possible to the participants and to control for baseline differences using pretest data (Wilkinson & APA Task Force on Statistical Inference, 1999).

When comparison groups cannot be obtained, comparison group scores can be simulated by modeling pretest developmental assessment scores as a function of chronological age and adding the expected change to the pretest score of each child to create an expected posttest score (McCall, Ryan, & Green, 1999). Observed posttest scores can then be contrasted to these expected scores. Suen, Bagnato, and Brickley (2000) have recently suggested a revised version of the constructed comparison that accounts for error in the slope estimate used to derive expected posttest scores.

**Using valid assessment measures.** Some evaluations used tests with little or no known reliability or validity, weakening the faith one can place in their results. The well-known Battelle Developmental Inventory (BDI; Newborg, Stock, Wnek, Guidubaldi, & Svinicki, 1984), for instance, has serious problems with the method for which standard scores are obtained, and these problems potentially can lead to erroneous findings in both clinical and research applications (Gilliam & Mayes, 2000). Kentucky, however, avoided that instrumentation problem by using intervention efficiency indexes (Bagnato & Neisworth, 1980), rather than standard scores in some of their analyses.<sup>9</sup> Nonetheless, evaluators should always use the most reliable and valid instruments available. If an evaluator must create a test specifically for an evaluation, complete evidence of its reliability and validity should be presented or reference provided for where that information can be obtained (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985).

**Avoiding abuses of null hypothesis significance testing (NHST).** Both Kentucky and South Carolina used random classmates as a comparison group for at least some analyses, attempting to infer program effects through similarities between participants and random classmates on various outcome measures. By failing to reject the null hypothesis, these evaluators sought to document that the program was successful at bridging the gap between these at-risk preschool participants and their more affluent classmates. Although this logic may seem to make intuitive sense, this type of analysis and interpretation represents a gross misapplication of NHST. By attempting to prove the null hypothesis by failure to reject it, these evaluators are in effect setting their study hypothesis alpha at .95, rather than the conventional .05, since any  $p$ -value from .05 to 1.00 would be interpreted as being close enough to complete equality of groups ( $p = 1.00$ ). In other words, the probability of finding the two groups to be similar by chance alone is 95%, which is not very compelling evidence for accepting the null hypothesis as true (Tukey, 1991). (See Meehl, 1967, for a discussion of the use of predicted values in NHST.) Additional analytic concerns in these state evaluations include the failure of some to use either inferential testing or standardized measures of effect size (instead relying on “eyeball” analyses of mean scores) and the frequent overuse of inferential testing without correcting for multiplicity. Since conducting numerous statistical analyses can increase significantly the Type I error rate (the likelihood of spurious positive findings), evaluations should use analytic procedures that reduce this likelihood (e.g., multivariate analyses with planned or post-hoc comparisons, such as

<sup>9</sup> The BDI has norm-reference groups so wide that standard scores can fluctuate by as much as two standard deviations depending on the exact age of the child. These fluctuations can result in significantly deflated standard scores when children become old enough to enter the next higher age band, potentially leading to erroneous findings when used in longitudinal studies. By focusing on age equivalence scores, rather than standard scores, analyses of BDI data using the intervention efficiency index can provide more reliable results.

the Scheffé analysis, or use one of the several Bonferroni-based correction procedures; Olejnik, Li, Supattathum, & Huberty, 1997).

***Sample size and the importance of effect size estimates.*** As previously presented, sample sizes varied greatly from state to state and had a profound effect on the results of NHST. Recall the case of school-administered academic achievement tests, where relatively small effects were found to be statistically significant when sample sizes were in the thousands, while effects about three times as large (and non-trivial by conventional definitions) were not found statistically significant when much smaller samples were used. Such instances expose the limitations of NHST and illustrate the need for standardized effect size estimates in order to assess these findings. Unfortunately, only two state evaluations provided effect size estimates in their presentation of results, with the remaining being computed by the present authors.

***Cumulative versus non-cumulative data.*** Both grade retention and special education placement were reported both cumulatively and non-cumulatively. When data were analyzed non-cumulatively at specific grade levels, results were mixed across grade levels. When data were analyzed cumulatively, however, results were always statistically significant for both grade retention and special education placement, and the standardized effects were of meaningful size. Arguably, variables where impacts are measured based on the occurrence or non-occurrence of a particular event are best analyzed cumulatively to account for differences in local policy or practice regarding when children are retained or referred for special education. A statistical alternative would be the use of survival analyses that address the amount of time before an event such as retention or special education placement occurs.

***Addressing bias in attrition.*** As previously stated, study attrition rates were found to approximate those of other program evaluations serving a similar population. However, no state evaluation provided results of tests for selective attrition. Given attrition rates ranging from 10% to 25% per year, it is important to assess the degree to which subjects lost to attrition differ from those retained. Indeed, it appears that there may have been at least one source of biased attrition at work in many of these evaluations. Follow-up evaluations were typically conducted at particular grade levels (e.g., kindergarten, first grade, etc.), rather than at particular post-intervention time intervals (e.g., 1-year follow-up, 2-year follow-up, etc.). Since a reduced incidence of grade retention was among the most robust positive impacts found, a greater number of comparison children are retained in grade and are subsequently lost to follow-up evaluations. Assuming that these children who are retained in grade would have been among the lowest scoring children on follow-up measures (e.g., test scores, grades), it follows that biased attrition may have actually elevated the mean scores of the comparison groups on follow-up measures, reducing the likelihood for sustained impacts. Such issues could contribute to a methodologically produced “fade-out” of effects (Barnett, 1992).

### ***What is the Role of Program Quality?***

Process evaluation measuring program implementation and quality should be an essential first step to program evaluation (Gilliam, 2000a); however, few states ever considered the issue of program quality and none used it to determine whether the programs were mature enough for evaluation (Campbell, 1987). The frequent omission of formal measurement of program implementation and quality is unfortunate. When an evaluation fails to demonstrate positive impacts, should one conclude that the program model was flawed or should one conclude that the program model was not implemented appropriately? When evaluation findings are disappointing, how can the program be improved? Without program quality data, outcome evaluations present

an incomplete picture and lead too frequently to findings that are difficult to interpret. This may be even truer when quality levels can be expected to be highly variable, as may be the case when state programs deliver their services through contractual agreements with a variety of local child care and education providers (Gilliam, 2000b).

Only three states completed analyses relating classroom quality indicators to program impacts. *South Carolina* found that the degree of teacher classroom management was positively related to the degree of program impacts on reading test scores in kindergarten. In fact, it was only after low scoring classrooms were removed from analyses that significant positive impacts were found (Appendix, Report # 42). *Michigan* also found a relationship between child developmental ratings in kindergarten and preschool program quality in the areas of program philosophy, use of funding, and administration and supervision (Appendix, Report # 4). In contrast to Michigan, South Carolina, and a host of other studies on the importance of quality (Bryant, Burchinal, Lau, & Sparling, 1994; CQO Study Team, 1995, 1999; Howes, Galinsky, & Shinn, 1998; Love, Schochet, & Meckstroth, 1996), *Kentucky* found no significant relationship between child outcomes and classroom quality, as measured by the *Early Childhood Environment Rating Scale* (Harms & Clifford, 1980). Examination of the quality scores, however, suggests that the variance may have been too restricted to find a significant relationship between program quality and child outcomes in this sample of 24 classrooms (Appendix, Report # 56).

## **CONCLUSIONS**

Impact evaluations of state-funded preschool programs vary considerably in their domains of interest, evaluation methodologies, and findings. Although only 13 state evaluations were summarized in this paper, a surprising amount of data were obtained given the many outcomes measured at various grade levels using multiple cohorts of children. Some evaluations consisted of as little as a pretest and posttest of participants, whereas others used comparison groups of varying resemblance to the participants. None used a randomly assigned control group. Some evaluations used relatively small samples of children, whereas others attempted to follow most, if not all, children enrolled in the program. These evaluations represent our current best estimate of the impact of this important and increasingly prevalent type of preschool program for mostly low-income children. Therefore, despite considerable differences and limitations in the methods used to evaluate these programs, a review of their findings is useful.

Considerably more needs to be known about the effectiveness of state-funded preschool programs. Although this paper may add somewhat to our understanding of the overall impact of these programs, it may add more to our understanding of the many ways in which they are currently being evaluated. Clearly, some evaluative methods stack the deck in favor of finding positive effects, while most appear to stack the deck against finding effects by such practices as using lower risk comparison groups, using sample sizes that lack sufficient power to detect non-trivial effects, and potentially allowing systematic attrition biases to mask potential impacts.

In this age of increased accountability, the effects of state-funded preschool programs, as well as the methods for evaluating them, will be important topics for scientific inquiry. Results from evaluations of state preschool programs may be used by policy-makers for informing decisions regarding the level of future funding for preschool programs for low-income children, as well as for addressing questions regarding who best should administer these programs (e.g., federal versus state versus local agencies). The answers to such questions have enormous implications for low-income children and families. Despite the methodological limitations

described earlier (limitations that may bias evaluations both for and against finding positive impacts), the findings are rather consistent in certain areas. Specifically, these state-funded preschool programs may help children enter school with a greater level of developmental competence, helping children to perform better in school during the critical early grades. These positive findings are encouraging for state-funded preschool programs, but, on the whole, appear to be no more or less encouraging than the findings for other large-scale preschool programs for low-income children, such as Head Start, which often suffer from similar methodological limitations in their evaluations (Barnett, 1995, 1998; McKey et al., 1985). Research over the past forty years has provided ample evidence that high-quality preschool programs can produce meaningful effects for low-income children. More attention, however, must be paid how to best achieve and sustain high-quality preschool services during broad implementation (such as with state- and federal-funded programs) and how to best evaluate these programs.

**Acknowledgement:** The authors wish to acknowledge the important contributions of Dr. Carol H. Ripple, whose work regarding state preschool programs has always been an important driving force for this paper. Data presented in this paper were first shared at the 1999 biennial meeting of the Society for Research in Child Development, where many helpful suggestions and critiques were offered by valued colleagues, in particular Dr. Samuel Meisels and Dr. Michael Lopez. We also thank Dr. Marion Hyson and the anonymous reviewers at *Early Childhood Research Quarterly*, for their highly useful suggestions during the review process.

## **APPENDIX: STATE PREKINDERGARTEN EVALUATION REPORTS REVIEWED**

1. *Experimental Prekindergarten Program: Follow-up Study*. (1985). Unpublished draft.
2. Flint, D. L., Hick, T. L., Irvine, D. J., Horan, M. D., & Kukuk, S. E. (1979). *Effects of exposure to prekindergarten on five social competency factors (Technical paper #8)*. Albany: New York State Education Department, Prekindergarten Evaluation Unit.
3. Flint, D. L., Hick, T. L., Kukuk, S. E., Horan, M. D., & Irvine, D. J. (1978). Effects of prekindergarten on the cognitive performance of children at the end of prekindergarten: Wave II. In D. J. Irvine (Chair), *Ongoing research on the New York State Experimental Prekindergarten Program*. Symposium conducted at the annual meeting of the Northeastern Educational Research Association, Ellenville, NY.
4. Florian, J. E., Schweinhart, L. J., & Epstein, A. S. (1997, September). *Early returns: First year report of the Michigan School-Readiness Program Evaluation*. Ypsilanti, MI: High/Scope Press. [Note: Although this report is not available on-line, a recent follow-up evaluation can be accessed at <http://www.highscope.org/MSRP/Support/Points%20of%20Light.PDF>]
5. Hick, T. L., Flint, D. L., Kukuk, S. E., Horan, M. D., & Irvine, D. J. (1978). Effects of prekindergarten on the cognitive performance of children in kindergarten. In D. J. Irvine (Chair), *Ongoing research on the New York State Experimental Prekindergarten Program*. Symposium conducted at the annual meeting of the Northeastern Educational Research Association, Ellenville, NY.
6. Hick, T. L., Irvine, D. J., Horan, M. D., Flint, D. L., Kukuk, S. E., & Fallon, E. (1979, August). *Effects of parent involvement in a prekindergarten program on children's cognitive performance*. Albany: New York State Education Department, Prekindergarten Evaluation Unit.
7. Hick, T. L., Kukuk, S. E., Horan, M. D., & Irvine, D. J. (1977). *Effects of prekindergarten on three measures of cognitive development (Technical paper #3)*. Albany: New York State Education Department, Prekindergarten Evaluation Unit.
8. Hick, T. L., Kukuk, S. E., Horan, M. D., Irvine, D. J., & Flint, D. L. (1978). *Effects of exposure to prekindergarten on the development of social competency in children (Technical paper #5)*. Albany: New York State Education Department, Prekindergarten Evaluation Unit.

9. Horan, M. D., Irvine, D. J., Flint, D. L., & Hick, T. L. (1980). A prekindergarten program: Policy implications of the research. *Education and Urban Society*, 12(2), 193-210.
10. Irvine, D. J., Flint, D. L., Hick, T. L., Horan, M. D., & Kukuk, S. E. (1981). *Effects of an early childhood education program on children's verbal knowledge and general reasoning ability in third grade (Technical paper #10)*. Albany: New York State Education Department, Prekindergarten Evaluation Unit.
11. Irvine, D. J., Horan, M. D., Flint, D. L., Hick, T. L., & Kukuk, S. E. (1978, January). *Obtaining control groups for evaluating the New York State Experimental Prekindergarten Program*. Albany: New York State Education Department, Prekindergarten Evaluation Unit.
12. King, F. J., Cappellini, C. H., & Gravens, L. (1995, August). *A longitudinal study of the Florida Prekindergarten Early Intervention Program: Part III*. Tallahassee: Florida Department of Education.
13. King, F. J., Cappellini, C. H., & Rohani, F. (1995, December). *A longitudinal study of the Florida Prekindergarten Early Intervention Program: Part IV*. Tallahassee: Florida Department of Education.
14. King, F. J., & Gravens, L. (1994, February). *A longitudinal study of the Florida Prekindergarten Early Intervention Program: Part II*. Tallahassee: Florida Department of Education.
15. King, F. J., Rohani, F., & Cappellini, C. H. (1991, November). *The third party evaluation report on the Prekindergarten Early Intervention Program, the Florida First Start Program, the Children's Early Investment Program, and the Community Resource Mother or Father Program*. Tallahassee: Florida Department of Education.
16. King, F. J., Rohani, F., & Cappellini, C. H. (1993). *A four-year longitudinal study of the Florida Prekindergarten Early Intervention Program*. Tallahassee: Florida Department of Education.
17. King, F. J., Rohani, F., & Morris, D. (1994, November). *Parental involvement in the Prekindergarten Early Intervention Program*. Tallahassee: Florida Department of Education.
18. Louisiana Board of Elementary and Secondary Education. (n.d.). Unpublished raw data.
19. Marcon, R. A. (1987, March). *Evaluation of educational programs serving four-year-olds*. Washington, DC: District of Columbia Public Schools.
20. Marcon, R. A. (1987, December). *Early learning and early identification study: 1986-1987*. Washington, DC: District of Columbia Public Schools.
21. Marcon, R. A. (1989, February). *Early learning and early identification study: 1987-1988*. Washington, DC: District of Columbia Public Schools.
22. Marcon, R. A. (1990, December). *Early learning and early identification: Final report of the three year longitudinal study*. Washington, DC: District of Columbia Public Schools.
23. Marcon, R. A. (1994, March). *Early learning and early identification follow-up study: Transition from the early to the later childhood grades: 1990-93*. Washington, DC: District of Columbia Public Schools.
24. Marcon, R. A. (1994). Doing the right thing for children: Linking research and policy reform in the District of Columbia Public Schools. *Young Children*, 50(1), 8-20.
25. Marcon, R. A. (1999). Differential impact of preschool models on development and early learning of inner-city children: A three cohort study. *Developmental Psychology*, 35, 358-375.
26. Maryland State Department of Education. (1991). *An analysis of the long-term effects of the Extended Elementary Education Prekindergarten Program*. Baltimore, MD: Author.
27. McKean, K. & Thistlethwaite, P. C. (1996, June). *Arkansas governor's commission early childhood study: Year one 1995-1996*.
28. Northwest Regional Educational Laboratory. (1989). *Tracking success for children and families: ECEAP longitudinal evaluation study year one report*. Portland, OR: Author.
29. Northwest Regional Educational Laboratory. (1991). *Tracking success for children and families: ECEAP longitudinal evaluation study year 2 technical report*. Portland, OR: Author.
30. Northwest Regional Educational Laboratory. (1992). *Tracking success for children and families: ECEAP longitudinal evaluation study year 3 technical report*. Portland, OR: Author.
31. Northwest Regional Educational Laboratory. (1993). *1992 ECEAP longitudinal study and annual report: An evaluation of child and family development through comprehensive preschool services*. Portland, OR: Author. (ERIC Document Reproduction Service No. ED 360 422)
32. Northwest Regional Educational Laboratory. (1994). *1993 ECEAP longitudinal study: Year 5 technical report: An evaluation of child and family development through comprehensive preschool services*. Portland, OR: Author.
33. Northwest Regional Educational Laboratory. (1995). *Early childhood education and assistance program: An investment in children and families. 1994 longitudinal study: Year 6*. Portland, OR: Author.
34. Northwest Regional Educational Laboratory. (1998a). *Early Childhood Education and Assistance Program: An*

- investment in children and families: 1995 longitudinal study: Year 7. Summary.* Portland, OR: Author.
35. Northwest Regional Educational Laboratory. (1998b). *Early Childhood Education and Assistance Program: An investment in children and families: 1995 longitudinal study: Year 7.* Portland, OR: Author.
  36. Pilcher, L. C., & Kaufman-McMurrain, M. (1994). *Georgia prekindergarten program evaluation* (Georgia Department of Education, Contract No. 950969). Atlanta: Georgia State University, Department of Early Childhood Education.
  37. Pilcher, L. C., & Kaufman-McMurrain, M. (1995). *The longitudinal study of Georgia's prekindergarten children and families: 1994-1995.* Atlanta: Georgia State University, Department of Early Childhood Education.
  38. Pilcher, L. C., & Kaufman-McMurrain, M. (1996). *The longitudinal study of Georgia's prekindergarten children and families: 1995-1996.* Atlanta: Georgia State University, Department of Early Childhood Education.
  39. Quay (Pilcher), L. C., & Kaufman-McMurrain, M. (1993). *Georgia prekindergarten program evaluation* (Georgia Department of Education, Contract No. 940996). Atlanta: Georgia State University, Department of Early Childhood Education.
  40. Quay (Pilcher), L. C., Kaufman-McMurrain, M., Steele, D. C., & Minore, D. A. (1997). *The longitudinal evaluation of Georgia's prekindergarten program.* Paper presented at the biennial meeting of the Society for Research in Child Development.
  41. South Carolina Department of Education. (1986, August). *Early childhood development programs: Half-day programs for four-year olds. 1983-84 school year: Large sample evaluation report.* Columbia: South Carolina Department of Education, Office of Research. (Available from the South Carolina State Library, State Documents)
  42. South Carolina Department of Education. (1987, October). *Executive summary of the evaluation of South Carolina half-day programs for four-year-olds.* Columbia: South Carolina Department of Education, Office of Research. (Available from the South Carolina State Library, State Documents)
  43. South Carolina Department of Education. (1988, October). *The longitudinal evaluation of early childhood education: Interim report. The half-day child development program for four-year-olds* (The Office of Research Report Series). Columbia: South Carolina Department of Education, Office of Research. (Available from the South Carolina State Library, State Documents)
  44. South Carolina Department of Education. (1990, July). *The longitudinal evaluation of early childhood education: Third interim report-revised. The half-day child development program for four-year-olds* (The Office of Research Report Series). Columbia: South Carolina Department of Education, Office of Research. (Available from the South Carolina State Library, State Documents)
  45. South Carolina Department of Education. (1990, December). *The longitudinal evaluation of early childhood education: Fourth interim report. The half-day child development program for four-year-olds* (The Office of Policy Research Report Series). Columbia: South Carolina Department of Education, Office of Policy Research. (Available from the South Carolina State Library, State Documents)
  46. Squires, J. H. (1995). *Vermont early education initiative child follow-through survey: Preliminary results: Kindergarten, first and second grades.* Montpelier: Vermont Department of Education.
  47. Texas Education Agency. (1992, September). *Texas evaluation study of prekindergarten programs: Preliminary findings* (Publication No. GE2-091-08). Austin, TX: Author.
  48. Texas Education Agency. (1993, May). *Texas evaluation study of prekindergarten programs: Interim report* (Publication No. GE3-410-09). Austin, TX: Author.
  49. Texas Education Agency. (1995, July). *Texas evaluation study of prekindergarten programs: Final report summary* (Publication No. GE5-170-01). Austin, TX: Author.
  50. Texas Education Agency. (1995, July). *Texas evaluation study of prekindergarten programs: Final report* (Publication No. GE5-170-02). Austin, TX: Author.
  51. University of Kentucky. (1992). *Third party evaluation: Kentucky Education Reform Act (KERA) Preschool Programs. Final report.* Lexington: University of Kentucky, College of Human Environmental Sciences and College of Education.
  52. University of Kentucky. (1993). *Third party evaluation: Kentucky Education Reform Act (KERA) Preschool Programs. Final report.* Lexington: University of Kentucky, College of Education and College of Human Environmental Sciences.
  53. University of Kentucky. (1994). *Third party evaluation of the Kentucky Education Reform Act Preschool Programs.* Lexington: University of Kentucky, College of Education and College of Human Environmental Sciences.

54. University of Kentucky. (1995). *Third party evaluation of the Kentucky Education Reform Act Preschool Programs*. Lexington: University of Kentucky, College of Education and College of Human Environmental Sciences.
55. University of Kentucky. (1996). *Third party evaluation of the Kentucky Education Reform Act Preschool Programs*. Lexington: University of Kentucky, College of Education and College of Human Environmental Sciences.
56. University of Kentucky. (1997). *Third party evaluation of the Kentucky Education Reform Act Preschool Programs*. Lexington: University of Kentucky, College of Education and College of Human Environmental Sciences.
57. University of the State of New York, & New York State Education Department. (1977). *Preliminary report of findings: Evaluation of the New York State Experimental Prekindergarten Program*. Albany: Authors.
58. University of the State of New York, & New York State Education Department. (1978). *Second-year findings, with special emphasis on cognitive outcomes: Evaluation of the New York State Experimental Prekindergarten Program*. Albany: Authors.
59. University of the State of New York, & New York State Education Department. (1982). *Evaluation of the New York State Experimental Prekindergarten Program: Final report*. Albany: Authors. (ERIC Document Reproduction Service No. ED 219 123)
60. Wode, J., Salehi, S., & Eckroade, G. (1992). *An analysis of the long-term effects of the EEEP Program: The secondary years*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

## **REFERENCES**

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Bagnato, S. J., & Neisworth, J. T. (1980). The intervention efficiency index: An approach to preschool program accountability. *Exceptional Children*, 46, 264-269.

Barnett, W. S. (1995). Long-term effects of early childhood programs on cognitive and school outcomes. *The Future of Children*, 5(3), 25-50.

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

Barnett, W. S. (1992). Benefits of compensatory preschool education. *Journal of Human Resources*, 27, 279-312.

Bredekamp, S., & Copple, C. (Eds.). (1997). *Developmentally appropriate practice in early childhood programs* (Rev. ed.). Washington, DC: National Association for the Education of Young Children.

Bryant, D. M., Burchinal, M., Lau, L. B., & Sparling, J. J. (1994). Family and classroom correlates of Head Start children's developmental outcomes. *Early Childhood Research Quarterly*, 9, 289-309.

Campbell, D. T. (1987). Problems for the experimenting society in the interface between evaluation and service providers. In S. L. Kagan, D. R. Powell, B. Wissbourd, & E. Zigler (Eds.), *America's family support programs* (pp. 345-351). New Haven, CT: Yale University Press.

Coats Human Services Reauthorization Act of 1998, Pub. L. No. 105-285, § 102 [online]. Available: <http://web.lexis-nexis.com/universe/> [Accessed April 21, 2000].

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 69, 145-153.

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2<sup>nd</sup> ed.). Hillsdale, NJ: Erlbaum.
- Cohen, J. (1994). The earth is round ( $p < .05$ ). *American Psychologist*, 49, 997-1003.
- Colorado Department of Education (n.d.). *Colorado Preschool Program: Child progress in years 1-3: Summer, 89, through summer, 92*. Denver: Author.
- Cooke, B. (1992, March). *Changing times, changing families. Minnesota Early Childhood Family Education parent outcome interview study*. St. Paul: Minnesota Department of Education.
- Council for Early Childhood Professional Recognition. (1996). *The child development associate assessment system and competency standards: Preschool caregivers in center-based programs*. Washington, DC: Author.
- CQO Study Team. (1995). *Cost, quality, and child outcomes in child care centers: Technical report*. University of Colorado, Denver.
- CQO Study Team. (1999, June). *The children of the cost, quality, and outcomes study go to school*. University of Colorado, Denver. Available (executive summary): <http://www.fpg.unc.edu/~ncedl/pages/cqes.htm> [Accessed April 21, 2000].
- Fielden, F., Smith, D. B., Soper-Hepp, E., McNulty, B., & Randall, W. T. (1994, February). *An analysis of three year trends in the Colorado Preschool Program: 1994 report to the Colorado Legislature*. Denver: Colorado Department of Education.
- Gilliam, W. S. (2000a). On over-generalizing from overly-simplistic evaluations of complex social programs: In further response to Goodson, Layzer, St.Pierre and Bernstein. *Early Childhood Research Quarterly*, 15, 66-71.
- Gilliam, W. S. (2000b). *The School Readiness Initiative in South-Central Connecticut: Classroom quality, teacher training, and service provision. Final report of findings for fiscal year 1999*. Unpublished report, Yale University Child Study Center, New Haven, CT.
- Gilliam, W. S., & Mayes, L. C. (2000). Developmental assessment of infants and toddlers. In C. H. Zeanah, Jr. (Ed.), *Handbook of Infant Mental Health* (2nd ed.; pp. 236-248). New York: Guilford.
- Gilliam, W. S., & Ripple, C. H. (in press). What can be learned from state-funded preschool initiatives?: A data-based approach to the Head Start devolution debate. In E. Zigler & S. J. Styfco (Eds.), *The Head Start debates (friendly and otherwise)*. New Haven, CT: Yale University Press.
- Gilliam, W. S., Ripple, C. H., Zigler, E. F., & Leiter, V. (2000). Evaluating child and family demonstration initiatives: Lessons from the Comprehensive Child Development Program. *Early Childhood Research Quarterly*, 15, 41-59.
- Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.
- Goals 2000: Educate America Act of 1994, Pub. L. No. 103-227, § 102 [on-line]. Available: <http://web.lexis-nexis.com/universe/> [Accessed April 21, 2000].
- Gomby, D. S. (1999). Understanding evaluations of home visitation programs. *The Future of Children*, 9(1), 27-43.
- Guralnick, M. J. (Ed.). (1997). *The effectiveness of early intervention*. Baltimore, MD: Brookes.
- Harms, T., & Clifford, R. M. (1980). *Early childhood environment rating scale*. New York: Teachers College Press.

- Head Start Bureau. (1999). *Head Start program regulations (45 CFR, Parts 1301-1311)*. Available: [http://www2.acf.dhhs.gov/programs/hsb/regs/rg\\_index.htm](http://www2.acf.dhhs.gov/programs/hsb/regs/rg_index.htm) [Accessed August 16, 2000].
- Heubert, J. P., & Hauser, R. M. (Eds.). (1999). *High-stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.
- Hofferth, S. L., West, J., Henke, R., & Kaufman, P. (1994). *Access to early childhood programs for children at risk*. (ERIC Reproduction Service No. ED 370 715). Washington, DC: National Center for Education Statistics.
- Hohmann, M., & Weikart, D. P. (1995). *Educating young children: Action learning practices for preschool and child care programs*. Ypsilanti, MI: High/Scope Press.
- Horan, M. D., Irvine, D. J., Flint, D. L., & Hick, T. L. (1980). A prekindergarten program: Policy implications of the research. *Education and Urban Society*, 12, 193-210.
- Howes, C., Galinsky, E., & Shinn, M. (1998). *The Florida Child Care Quality Improvement Study: 1996 report*. New York: Families and Work Institute.
- Kaplan, J. (1998). State-funded prekindergarten programs. *Welfare Information Network*, 2(8), 1-11 [on-line]. Available: <http://www.welfareinfo.org/preschoo.htm> [Accessed April 21, 2000].
- Karoly, L. A., Greenwood, P. W., Everingham, S. S., Houbé, J., Kilburn, M. R., Rydell, C. P., Sanders, M., & Chiesa, J. (1998). *Investing in our children: What we know and don't know about the costs and benefits of early childhood interventions*. Washington, DC: Rand Corporation.
- Knitzer, J., & Page, S. (1998). *Map and track: State initiatives for young children and families*. New York: National Center for Children in Poverty.
- Lally, J. R., Mangione, P. L., & Honig, A. S. (1988). The Syracuse University Family Development Research Project: Long-range impact of an early intervention with low-income children and their families. In D. R. Powell (Ed.), *Parent education as early childhood intervention: Emerging directions in theory, research and practice* (pp. 79-104). Norwood, NJ: Ablex.
- Love, J. M., Schochet, P. Z., & Meckstroth, A. L. (1996). Are they in any real danger? What research does and does not tell us about childcare quality and children's well-being. In *Child Care Research and Policy Papers*. Princeton, NJ: Mathematica.
- Marcon, R. A. (1999). Differential impact of preschool models on development and early learning of inner-city children: A three-cohort study. *Developmental Psychology*, 35, 358-375.
- McCall, R., Larsen, L., & Ingram, A. (2000). *The science and policies of early childhood education and family services*. Unpublished manuscript, University of Pittsburgh, PA.
- McCall, R. B., Ryan, C. S., & Green, B. L. (1999). Some non-randomized constructed comparison groups for evaluating age-related outcomes of intervention programs. *American Journal of Evaluation*, 2, 213-226.
- McCartney, K., & Rosenthal, R. (2000). Effect size, practical importance, and social policy for children. *Child Development*, 71, 173-180.
- McKey, R., Condelli, L., Ganson, H., Barrett, B., McConkey, C., & Plantz, M. (1985). *The impact of Head Start on children, families, and communities: Final report of the Head Start Evaluation, Synthesis, and Utilization Project*. Washington, DC: U.S. Department of Health and Human Services.
- Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34, 103-115.

Mueller, M. R. (1996, April). *Immediate outcomes of lower-income participants in Minnesota's Universal Access Early Childhood Family Education*. St. Paul: Minnesota Department of Children, Families and Learning.

National Association for the Education of Young Children. (1998). *Accreditation criteria and procedures of the National Association for the Education of Young Children: 1998 edition*. Washington, DC: Author.

National Education Goals Panel. (1996). *The National Education Goals Report: Building a nation of learners* (GPO No. 1996-415-143/60328). Washington, DC: U.S. Department of Education.

Newborg, J., Stock, J. R., Wnek, L., Guidubaldi, J., & Svinicki, J. (1984). *Battelle developmental inventory: Examiner's manual*. Allen, TX: DLM/Teaching Resources.

Olejnik, S., Li, J., Supattathum, S., & Huberty, C. J. (1997). Multiple testing and statistical power with modified Bonferroni procedures. *Journal of Educational and Behavioral Statistics*, 22, 389-406.

Phillips, V., Boysen, T. C., & Schuster, S. A. (1997). Psychology's role in statewide education reform: Kentucky as an example. *American Psychologist*, 52, 250-255.

Ripple, C. H., Gilliam, W. S., Chanana, N., & Zigler, E. (1999). Will fifty cooks spoil the broth? The debate over entrusting Head Start to the states. *American Psychologist*, 54, 327-343.

Rosenthal, R. (1993). Cumulating evidence. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 519-559). Hillsdale, NJ: Lawrence Erlbaum.

Sawhill, I. V. (1999, Fall). Kids need an early start: Universal preschool education may be the best investment Americans can make in our children's education – and our nation's future. *Blueprint Magazine*. Available: <http://www.brook.edu/views/articles/sawhill/19990825.htm> [Accessed April 21, 2000].

Schulman, K., Blank, H., & Ewen, D. (1999). *Seeds of success: State prekindergarten initiatives 1998-1999*. Washington, DC: Children's Defense Fund.

Schweinhart, L. J., Barnes, H. V., & Weikart, D. P. (1993). Significant benefits: The High/Scope Perry Preschool study through age 27. *Monographs of the High/Scope Educational Research Foundation*, 10. Ypsilanti, MI: High/Scope Educational Research Foundation.

Smith, S. L., Fairchild, M., & Groginsky, S. (1997). *Early childhood care and education: An investment that works* (2nd ed.). Washington, DC: National Conference of State Legislatures.

Suen, H., Bagnato, S. J., & Brickley, D. (2000). *Developmental outcome and impact of the Early Childhood Initiative (ECI) Model: Statistical analysis of the first 16 months*. Pittsburgh, PA: Children's Hospital of Pittsburgh, The UCLID Center at the University of Pittsburgh.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science*, 6, 100-116.

United States General Accounting Office. (1999, November). *Education and care: Early childhood programs and services for low-income families* (GAO Report No. B-281005). Washington, DC: Author.

University of the State of New York, & New York State Education Department. (1977). *Preliminary report of findings: Evaluation of the New York State Experimental Prekindergarten Program*. Albany: Authors.

West, J., Hausken, E. G., & Collins, M. (1993). *Profile of preschool children's child care and early education program participation: National Household Education Survey* (NCES 93-

133). Washington, DC: National Center for Education Statistics, Office of Educational Research and Improvement, U.S. Department of Education.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604.

Yoshikawa, H. (1995). Long-term effects of early childhood programs on social outcomes and delinquency. *The Future of Children*, 5(3), 51-75.

Young, K. T., Marsland, K. W., & Zigler, E. (1997). The regulatory status of center-based infant and toddler child care. *American Journal of Orthopsychiatry*, 67, 535-544.

Zigler, E. (1998). By what goals should Head Start be assessed? *Children's Services: Social Policy, Research, and Practice*, 1, 5-18.

Zigler, E., Taussig, C., & Black, K. (1992). Early childhood intervention: A promising preventative for juvenile delinquency. *American Psychologist*, 47, 997-1006.