Early Education Policy Alternatives: Comparing Quality and Outcomes of Head Start and State Prekindergarten

Gary T. Henry, Craig S. Gordon and Dana K. Rickman

EDUCATIONAL EVALUATION AND POLICY ANALYSIS 2006; 28; 77
DOI: 10.3102/01623737028001077

Educational Evaluation and Policy Analysis
Spring 2006, Vol. 28, No. 1, pp. 77–99

Early Education Policy Alternatives: Comparing Quality and Outcomes of Head Start and State Prekindergarten

Gary T. Henry, Craig S. Gordon, and Dana K. Rickman
Andrew Young School of Policy Studies, Georgia State University

The debates over the 2003 reauthorization of Head Start highlighted a controversy about the devolution of federal early education policy. At the center of the debate is the concern that state control of early education programs will reduce the quality and effectiveness of federal support for children living in poverty, and their families. The current fragmentation of early education policy, with both federal Head Start programs and state-subsidized prekindergarten programs operating in close proximity, presents an opportunity to compare the programs' quality and effectiveness within a region of common support. In this study, propensity score techniques were used to match a probability sample of Head Start participants in Georgia with a group of children who were eligible for Head Start but who attended the state prekindergarten program in Georgia. The two groups were statistically similar at the beginning of their preschool year on three of four direct assessments ($p < .05$), but by the beginning of kindergarten the children attending the state prekindergarten program posted higher developmental outcomes on five of six direct assessments ($p < .05$) and 14 of 17 ratings by kindergarten teachers ($p < .05$). This study indicates that economically disadvantaged children who attended Georgia's universal prekindergarten entered kindergarten at least as well prepared as similar children who attended the Head Start program.

Keywords: *education policy, universal kindergarten*

It is doubtful that any policy action will have more influence on the development and life opportunities of young children living in high-poverty families in the United States than the pending reauthorization of Head Start. The 1994 reauthorization of Head Start set in motion several programmatic changes, including clarifying the focus on school readiness and establishing performance measures for Head Start to deliver higher quality services and employ more educated teachers. Since that reauthorization, expenditures have doubled to $6.7 billion in 2003, and the number of children receiving services increased by 22% to 909,608 children during the same period (Head Start Bureau, 2004). While funding increases of this magnitude are unlikely over the next few years, there is the possibility of significant policy change in Head Start, the nation's largest early education program.

Among the issues that have risen to the top of the agenda for the current reauthorization is a contentious policy debate about devolving control of Head Start to the states. Head Start is the cornerstone of the federal policies supporting early childhood development and education, and is one of the few remaining antipoverty programs initiated by President Lyndon Johnson. Currently, the program bypasses the states and directly funds

An earlier version of this article was presented at the Fall Research Conference of the American Association for Policy Analysis and Management, October 28–30, 2004, in Atlanta, Georgia. This research was funded in part by the National Institute for Early Education Research, which is generously funded by the Pew Charitable Trust, and the Georgia Department of Early Care and Learning. The content of the article and the interpretations of the findings are the responsibility of the authors.

independent local agencies (grantees), which operate comprehensive school readiness and social support programs targeted primarily to 3- and 4-year-olds from economically disadvantaged homes, and their families. The issue of devolution has cast a long shadow over the reauthorization debate, challenging longstanding political commitments and administrative relationships as well as raising questions about the consequences of such a change. At issue is not the lack of state experience with early childhood policies—42 states have substantial experience offering early childhood education programs (Quality Counts, 2002)—but the lack of evidence about the programs' quality, effectiveness, and commitment to serve families considered "the poorest of the poor."

Supported at least in part by the public's eroding confidence in the federal government (Shaw & Reinhart, 2001), devolution has been a popular reform strategy at the federal level. The passage of the Personal Responsibility and Work Opportunity Reconciliation Act (PRWORA) in 1995 serves as a recent precedent for devolving federal programs to the states. In the case of welfare reform, states could apply for and receive waivers granting exemptions from federal regulations in order to offer alternative welfare programs, most of which required welfare recipients to work whenever possible. The federal requirement that states granted waivers undertake randomized experiments to evaluate their alternative policies played a major role in producing the bipartisan support necessary to pass PRWORA (Greenburg, Mandell, & Onstott, 2000). What must be noted is that the proposals to devolve welfare completely to the states and require each state to implement "work-first" reforms, was accompanied by head-to-head comparisons of the existing welfare system and the work-first reforms in the states that had received waivers. In contrast, the debates concerning the devolution of Head Start to the states have not been informed by directly comparable evidence about the quality and effects of the existing Head Start program and alternative state prekindergarten programs.

The dearth of evidence about the impacts of the Head Start program complicates the issue of devolution of Head Start. As early as 1997, evaluating Head Start's effectiveness had been placed on the agenda by the U.S. General Accounting Office (GAO) (1997). After reviewing approximately 600 articles and manuscripts, the GAO concluded that "[t]he body of research is inadequate for use in drawing conclusions about the impact of the national program in any area in which Head Start provides services such as school readiness or health-related services" (p. 2). Furthermore, the agency stated that planned research funded by the Department of Health and Human Services "will provide little information on the impact of regular Head Start programs" (p. 2). Since then a randomized experiment assessing Head Start's impacts has been initiated (Office of Planning Research and Evaluation, 2003).

The lack of evidence concerning the effectiveness of state early education programs has been at issue as well. A meta-evaluation of state prekindergarten programs highlighted substantial issues in research designs, measures, and analytical methods of the evaluations conducted prior to the review, and the lack of evidence about the implementations and outcomes of the state programs (Gilliam & Zigler, 2001). However, the body of research has grown since this meta-evaluation, and the research shows that state prekindergarten programs can produce positive effects on short-term measures (Gormley & Gayer, 2005); Gormley, Gayer, Phillips, & Dawson, 2005; Henry et al., 2003) and are associated with higher levels of achievement in later years (Grissmer, Flanagan, Kawata, & Williamson, 2000).

In this study, we begin to address the lack of directly comparable information on the quality and outcomes of Head Start and state prekindergarten programs. We assess children's developmental outcomes, including both academic skill development (cognitive and language skills) and broader social outcomes (social skills, health and well-being, and overall school readiness), along with program quality. We acknowledge that Head Start has additional goals related to parent involvement and parental human resource development that are not included in this study. We have chosen to focus on a comprehensive list of children's developmental outcomes that reflect the highest priorities for state prekindergarten programs and, since the two most recent reauthorizations, for Head Start. Moreover skills such as those included in this study have been shown to relate directly to children's later success in school (Reynolds, 2000).

Proponents of each of the two policy alternatives offer different hypotheses about the expected effectiveness of the policies. Head Start provides a comprehensive collection of services, including health and parental outreach, which may lead to better developmental outcomes for participants by addressing or preventing ill health and engaging parents more fully in their children's education. In contrast, state prekindergarten programs can be developed and administered by state and local agencies that could be more aware of the needs within the communities they serve and, in turn, produce better developmental outcomes. In addition, peer influences—that is, being in classrooms with peers who have more developed language, cognitive, and social skills—could produce positive skills gains for other children in those classrooms (Henry & Rickman, 2006). There is likely to be greater opportunity for positive peer effects in universal prekindergarten classes, because eligibility is not highly correlated with the children's skill levels and because these state programs are less likely to allow younger children in classrooms with the prekindergarten students.

In this study, we address the following research questions:

1. Does the quality of services and level of teacher education differ between Head Start and a state prekindergarten program?
2. Do children's developmental outcomes, including language skills, cognitive skills, social skills, and school readiness, differ between children from high-poverty households who receive services in Head Start and those who receive services from a state prekindergarten program?

We compare program quality and the developmental outcomes at entry into kindergarten for a multistage probability sample of children who participated in Head Start in Georgia who were matched with a sample of children from economically disadvantaged families who participated in Georgia's universal prekindergarten program (Pre-K). To compare these two groups directly, we use propensity scoring to minimize bias between the samples of children and propensity weights to adjust the means for each group. Propensity score matching makes the appropriately modeled outcomes independent of assignment to treatment. We compare the group differences for three types of measures: (1) mean standardized test scores for four standardized test scores at three points: entry into preschool, end of preschool, and start of kindergarten; (2) means of children's skills, readiness for school, and health and well-being as rated by their kindergarten teachers; and (3) means of three measures of the quality of each program. The comparison of the test scores at entry into preschool provides an independent assessment of the differences between the two groups at the beginning of the study. Our overall goal for the study is to provide a direct comparison of the program quality and developmental status of two similar groups of young children, one of whom had participated in Head Start and the other in a state prekindergarten program when they entered kindergarten.

Differences in Early Education Programs

A body of evidence has substantiated that high-quality early education programs can be both successful in improving children's developmental status (Consortium for Longitudinal Studies, 1983; Schweinhart & Weikart, 1997) and cost-effective (Barnett, 1991). Studies of large-scale public preschool programs indicate that the programs can contribute to increased development, overall school readiness, and future success of the children who participate in them (Garces, Thomas, & Currie, 2002; Gormley & Gayer, 2005; Gormley et al., 2005; Henry et al., 2003; Reynolds, 2000; Reynolds, Temple, Robertson, & Mann, 2001). Additionally, state expenditures for prekindergarten programs are associated with higher levels of educational performance on the state-level assessments in reading and math administered by the U.S. Department of Education and known as the National Assessment of Educational Progress (Grissmer, Flanagan, Kawata, & Williamson, 2000). Finally, the quality with which the services are implemented in early care and education environments is related to improved student success (NICHD Early Childcare Research Network, 2002; NICHD Early Childcare Research Network & Duncan, 2003; Peisner-Feinberg & Burchinal, 1997). What is absent in the current research literature, and from the federal early education policy debate specifically, is a comparative assessment of the outcomes and quality of the two most prominent early education policy models: Head Start and state prekindergarten programs. We begin with a description of these programs, including eligibility for services, size of the program, funding of the program for FY 2003, administration and operation of sites, goals and objectives, and teacher credential requirements. In addition, we provide information comparing Georgia Pre-K to other state prekindergarten programs and information comparing Head Start in the southern United States to the rest of the nation.

Georgia Pre-K

The nation's first universal prekindergarten program, Georgia Pre-K is open to all 4-year-old children residing in the state whose parents choose to enroll them, regardless of household means. In 1996–1997, the program served more than 57,000 4-year-olds. In 2001–2002, the year in which the sample for this study was selected, the program had expanded to serve 63,613 children, 25,711 (40%) of whom were classified as at-risk based on indicators of family income (Georgia Office of Educational Accountability, 2003). In Georgia, 38% (825,824) of children live in low-income families (The National Center for Children in Poverty, 2004).[1] The state expended approximately $216.3 million to operate the program in FY 2003.

The program is administered by Bright From the Start: Georgia Department of Early Care and Learning (DECAL).[2] The providers are local public schools (42%), not-for-profit organizations (12%), or private for-profit firms (46%) (Georgia Office of School Readiness, 2003). The Georgia Pre-K program's primary goal is preparing children for success in school. The program has established a comprehensive set of learning goals including language and communication skills as well as social and emotional development. To qualify as a lead teacher, professional staff must have at least a technical school diploma or 2-year college degree in a field directly related to early education or child development. Roughly 80% of lead teachers have a college degree in a field related to child development, family studies, or early education (Henry et al., 2004). Each classroom can enroll up to 20 students and must have a lead teacher and teacher's aide in the classroom whenever the children are present. In exchange for a flat payment per student from DECAL, providers must agree to offer full-day services (at least 6.5 hours) that follow the local school calendar (minimum of 180 days per year). However, the flat payment, which ranges from $2,200 to $3,475 per student, varies slightly based on program location and lead teacher credentials.[3]

Comparative data on state prekindergarten programs indicate that Georgia's Pre-K program has achieved a high degree of access (ranked second in the nation), is of average quality, and provides resources on par with the other state programs (Barnett, Hustedt, Robin, & Schulman, 2004). For example, the National Institute of Early Education Research (NIEER) reports the Georgia Pre-K program met six of 10 quality standards, exactly equal to the median for the 44 state programs in the 38 states rated (Barnett et al., 2004). Georgia did not meet the NIEER standards for provision of comprehensive services, education requirements for lead teachers or assistant teachers, or in-service training standard for lead teachers (Barnett et al., 2004). Per-pupil funding for Georgia's Pre-K program ranked 12th among the states at $3,824 for FY 2003 (Barnett et al., 2004). These data suggest that Georgia's program, neither at the top or nor the bottom of state programs, provides a fair basis of comparison with Head Start.

Head Start

Head Start is a national program that provides comprehensive, developmental services for low-income preschool children and their families. Eligibility for services is based on family income (below the federal poverty line of $18,400 for a family of four), receipt of public assistance, or having a child in foster care (Hart & Schumacher, 2004). In Georgia, Head Start serves nearly 20,000 children ranging from 3 to 5 years old, in 33 different programs covering 157 of Georgia's 159 counties (Georgia Head Start Collaboration Office, 2003), at an average program expenditure of $6,998 per child in FY 03 (Head Start, 2004). The total expenditure for Head Start in Georgia for FY 2003 was $163.8 million (Head Start, 2004). For the 2001–2002 school year, Head Start provided spaces for approximately 10,976 4-year-olds in Georgia (Georgia Head Start Collaboration Office, 2003). Funded at approximately $6.7 billion nationally, Head Start serves more than 900,000 children and their families each year (Butler & Gish, 2003). Nationally, it is estimated that Head Start serves roughly 50% of children eligible for the services at any given time (Barnett et al., 2004).

The Head Start program is administered by the Head Start Bureau within the Administration for Children and Families in the U.S. Department of Health and Human Services. The providers of Head Start services are federal grantees and include not-for-profit organizations, local school systems, and community organizations. Nationally, Head Start has five objectives on which their performance measures are based: (1) enhance children's growth and development; (2) strengthen families as the primary nurturers of their children; (3) provide children with educational, health, and nutritional services; (4) link children and families to needed community services; and (5) ensure well-managed programs that involve parents in decision-making (Zill et al., 1998). In recent years, Head Start goals have focused increasingly on academic skills related to school readiness, such as comprehension of spoken English, vocabulary, letter naming, phonological awareness, and early math (Zill et al., 1998). The program in Georgia is designed to address developmental goals for children, employment and self-sufficiency goals for adults, and support for parents in their work and in their roles as parents (Georgia Office of School Readiness, 2003).

One Head Start performance objective is that half its classroom teachers in center-based programs have an associate, baccalaureate, or advanced degree in early childhood education or a degree in a related field, with preschool teaching experience. In center-based programs, which were the exclusive focus of this study, the teacher must have a Child Development Associate (CDA) credential, its state-level equivalent, or meet the college degree and experience requirements (Advisory Committee on Head Start Research and Evaluation, 1999). Head Start programs in Georgia included in this study offered center-based services for at least 6 hours per day and for the part of the year that is reasonably coterminus with the local school calendar.

The children enrolled in Head Start programs across the south (including Georgia) differ from children in Head Start programs across the nation. A descriptive report of Head Start families in the FACES Study (O'Brien et al., 2002) indicates that children in the southern region faced greater risk factors than their counterparts in other regions. For example, Head Start children living in the south are less likely to have regular health-care coverage than Head Start children elsewhere. The primary caregiver of a Head Start child living in the south is more likely to be under the age of 29, have a lower income level, and have a lower education level than primary caregivers of Head Start children living in other parts of the nation.

Design, Sample, Measures, and Data

Design

The rational decision-making model prescribes comparing alternative policies on a common set of measures (Bardach, 2002). When the goal is evidence to influence attitudes and actions related to policy reform, the optimal research design is considered to be an experiment with random assignment of members of the target population to alternative polices, such as that used in the welfare-waiver experimental studies. However, randomized experiments have been underrepresented in evaluation of educational policies for many reasons (Cook, 2002). In circumstances where evidence from randomized experiments is unavailable, research has often relied on quasi-experimental designs, one of the strongest of which is a matched-sample design using propensity score matching.

Randomized designs and quasi-experiments using propensity score matching have a common and very desirable characteristic: the assignment to treatment is independent of the observed covariates. Random assignment also removes bias stemming from unobserved characteristics, and randomized designs are widely considered to provide unbiased estimates of the differences between two groups (although bias can arise even when assignment to treatment is randomized) (Heckman, Ichimura, Smith, & Todd, 1998; Manski & Garfinkle, 1992). Independence of the observed covariates is an important characteristic of propensity score matching designs that represents an improvement over regression-based controls in that it eliminates a major source of bias in regression which is known as "selection on observables" bias (Ravallion, 2001) and avoids results that have been extrapolated outside of the region of data defined by the two groups being compared (Cochran, 1965). However, neither random assignment nor propensity score matching assures the equivalence of the groups being compared, although when the samples of treated and control (or alternative treatment) groups are large enough, randomization generally yields equivalent groups. Design-based approaches to equivalence, such as random assignment, allow for the formation of groups without access to outcome data. This is a desirable attribute of any study (Rubin, 2001) and can be a characteristic of propensity score matching as well, and it was in this study.

In this study, we used three analyses to assess the equivalence of the Head State and state Pre-K groups and report the results of those later. As with random assignment studies, large samples are likely to yield closer approximations of pretreatment equivalence. Although sample size was a limiting design attribute in this study, the results of the three analyses show that bias was substantially reduced but not entirely eliminated by the propensity score matching.

The propensity score matching approach used in this study involved four steps: (1) select a probability sample of 4-year-old children attending Head Start in Georgia; (2) select a group of children who attended Georgia's Pre-K program but who were potentially eligible for Head Start as matches for each child in the Head Start sample using propensity score analysis; (3) develop weights to adjust for the likelihood that a child would have participated in Head Start; and (4) model differences in program quality and children's skills and readiness measures at the beginning of preschool, at the end of preschool, and at the beginning of kindergarten using the weights to adjust for differences in the samples of children attending Head Start and Pre-K. The method ensures independence of assignment to treatment across the observed covariates. However, the process of selecting the two groups to be compared requires several decisions that can make the groups more or less equivalent and the comparison of outcomes more or less generalizable. Details of the sampling, measures, formation of the matched Pre-K comparison group, and the weights are used to model program quality and children's outcomes.

Sample

A probability sample of 4-year-olds receiving early education services under the auspices of Head Start and the Georgia Pre-K Program was selected.[4] First, the counties in Georgia were stratified by the estimated number of 4-year-olds living in the county. In the first stage of the sample, counties were selected from each stratum. In the second stage, sites from the selected counties were selected from the lists of Head Start sites and Pre-K sites provided by their respective administrative agencies. One class was selected at random from sites that offered more than one class. Finally, five 4-year-old children were randomly selected from each class to participate in the study. The sampling procedures not only ensured that the Head Start sample was a probability sample of Head Start children in the state, the selection of children in the two groups from the same 24 counties within Georgia ensured that the samples were drawn from a region of common local support, which can mitigate a major source of bias (Heckman et al., 1998).

Ninety-eight Pre-K and Head Start sites were chosen through the stratified random sampling procedure and all agreed to participate, which was quite remarkable when compared with other studies of early education programs. We sampled children after obtaining parental consent (75% or more consented in most sites), permitting us to collect data on 353 Pre-K children and 134 Head Start participants. The number of Head Start children available for matching was reduced to 114 because four children moved out of state and were unavailable for testing; nine children withdrew from the Head Start program during 2001–2002 and did not receive a full year of service; and seven children did not have valid baseline measures from fall 2001. To enhance the generalizability of the comparison of developmental outcomes, we attempted to match as many of the Head Start children with children attending Georgia Pre-K, to maintain the comparison with a probability sample of children attending Head Start in Georgia.

Our design objective was to match the probability sample of 114 Head Start children to a group of children who would have been eligible to participate in Head Start but who, for any of a number of reasons, attended the Georgia Pre-K program. All 4-year-old children enrolled in Head Start in Georgia would have been eligible to participate in the Georgia Pre-K program. However, the participation of Georgia Pre-K children in Head Start would have been limited to those who met the federal eligibility requirements. The comparison group of children participating in Georgia's Pre-K Program was drawn by selecting a subsample of economically disadvantaged children who attended Pre-K, but who could have met the federal Head Start eligibility requirements. Of the original 353 Pre-K children, we classified 201 children as being potentially eligible for Head Start.[5] We made this classification based on administrative data from Temporary Assistance to Needy Families (TANF); Pre-K enrollment records indicating eligibility for food stamps, Supplemental Security Income, Medicaid, TANF, the Child Care and Parent Services program, or PeachCare for Kids; participation in foster care, or a reported family income of less than $35,000, which could make them eligible for assistance. These 201 represented the pool from the Georgia Pre-K program that was eligible for matching with the probability sample of Head Start children.

Data Collection

We collected data on four of the dimensions of children's developmental outcomes recommended by the National Education Goals Panel on School Readiness (Kagan, Moore, & Bradekamp, 1995). In addition, we collected data on child and family characteristics, teaching practices, teacher attitudes, and classroom quality. The study used a combination of 13 instruments, including direct assessments (standardized and nonstandardized), teacher surveys, teacher rating forms, parent surveys, and instruments for directly observing classroom quality and teacher–child interactions. The direct assessments produced standardized (norm-referenced) scores for receptive vocabulary, letter and word recognition, expressive language, phonological processing, and cognitive skills, in addition to scores on basic skills, such as naming numbers and colors. Kindergarten teachers were asked to rate other dimensions of the children's development, including social behaviors and health and well-being. Finally, child and family characteristics were obtained through parent surveys and administrative data. Response rates varied: 70% for the parent survey administered during the fall of the children's preschool year, 75% for teacher surveys, and 86% for direct assessments at the end of preschool and beginning of kindergarten.

Direct assessments of children

One study objective was to measure the developmental status of 4-year-olds participating in Georgia Pre-K and Head Start as comprehensively and accurately as possible without overburdening the children and schools. The children were directly assessed by trained assessors in the fall and spring of their preschool year and fall of their kindergarten year. Pre-K and Head Start children were tested during the same period, and all test scores were standardized.[6] These assessments measured: (1) cognition [Applied Problems—Woodcock Johnson Test Of Achievement III (WJIII)] and (2) language development [Peabody Picture Vocabulary Test (PPVT)]; Letter-Word Recognition (WJIII); Oral and Written Language Scales (OWLS); Sound Matching [Comprehensive Test of Phonological Processing (CTOPS)]; Elision (CTOPS). In this article, we report on the six standardized direct assessments administered by trained professionals listed in Table 1.

Teacher ratings

In addition to the direct assessments of skills, we collected teacher ratings on developmentally related outcomes including academic skills, social skills, health and well-being, communication skills, and general readiness. During fall 2002, we asked kindergarten teachers to rate all the children

TABLE 1
Georgia Early Childhood Developmental Assessments Instruments

Developmental area	Instrument
Cognition	Woodcock Johnson Test of Achievement-III (Woodcock, McGrew, & Mather, 2001, Applied Problems subtest)
	Receptive language (vocabulary): Peabody Picture Vocabulary Test-III, Form A (Dunn, Dunn, & Dunn, 1997)
	Recognition of letters and words: Woodcock Johnson Test of Achievement-III (Woodcock et al., 2001, Letter-Word Identification subtest)
Language development	Expressive language: Oral and Written Language Scales (Carrow-Woolfolk, 1995, Expressive subtest)
	Sound matching: Comprehensive Test of Phonological Processing (Wagner, Torgesen, & Rashotte, 1999)
	Elision: Comprehensive Test of Phonological Processing (Wagner, Torgesen, & Rashotte, 1999)

in the study on a number of dimensions, including academics, health and well-being, creativity, communication skills, behavior, and kindergarten readiness. We employed kindergarten teacher ratings for this study because the kindergarten teachers are more likely to see students using a greater range of skills and because nearly all the kindergarten teachers have a college education. Mashburn and Henry (2004) have shown that college-educated teachers are more consistent and reliable raters. Teachers' ratings were measured on a 7-point scale in which 4 indicates an average rating and 7 indicates an extraordinarily good rating (for an assessment of the validity and reliability of rating instrument used see Mashburn and Henry, 2004).

Parental and teacher surveys

Surveys were used to collect data about children, family, and classroom characteristics. Parents and teachers were surveyed about children's characteristics such as age, sex, and race. Families were surveyed at the beginning and end of their children's prekindergarten year and during the fall of their kindergarten year. When multiple responses were received, the most recent response was used. These surveys collected a comprehensive set of data about the children's families, including parental education, receipt of means-tested benefits, income, age of mother, marital status, parental employment, and information about health and wellness screenings prior to preschool. Teachers provided data on classroom composition, their own educational attainment and teaching credentials, and other attributes related to their teaching and the classroom.

Observations of classrooms

The Early Childhood Environmental Rating Scale Revised (ECERS-R), a directly observed, standard measure of quality used frequently in studies of the effects of center-based childcare and preschool, was used to rate the quality of the classroom environment in each study site. Trained raters conducted full-day classroom observations in all 93 preschool classrooms included in this study using the ECERS-R to measure the interactions, resources, and climate of the children's learning environment. The items rated include structural components of the classroom and the interactions between and among the children and teachers. Ratings were conducted during the late winter and early spring 2002 to avoid contamination of the observations during preschool start-up, end, or winter holidays.[7]

Administrative and other extant data

We used administrative databases to supplement the parental surveys for measures of participation in federal poverty programs. Two TANF databases were used to identify children whose families were receiving assistance. Families were coded as having received TANF if: (1) parents reported TANF receipt on the parent survey; (2) they were found in the active TANF database in December 2003; or (3) they were found in the active TANF database in March 2001.[8] We also collected data from administrative sources on the number of Head Start and Pre-K spaces for 4-year-olds that were available in each county by the type of provider offering the spaces. Finally, we collected U.S. Census data on counties for use in the propensity score matching.

Missing data

To correct for incomplete records or survey responses, we imputed data using a multiple imputation method (Little & Rubin, 1987; Schafer, 1997; Schafer & Graham, 2002).[9] Ten data sets were imputed, and each was analyzed separately. Each of the 10 test scores and their bootstrap standard errors were averaged for each period for which data were available and, along with the average differences, were reported in the tables. Little to no data (less than 5%) were missing for: (1) demographic data; (2) preschool test scores; (3) classroom characteristics; (4) county data; and (5) ECERS-R. Substantially more data were missing on: (1) some personal financial information (20%–40% missing); (2) kindergarten test scores (15%–20% missing); and (3) household characteristics, such as the number of children in the home (20%–30% missing). In addition to all the outcome variables and the variables used in the propensity score model (see Appendix A), we included the following measures in the multiple imputation procedures: (1) Pre-K and Head Start slots per capita in each county; (2) additional county-level census data; (3) summer and after-school program availability; and (4) additional family characteristics.

Propensity score matching

One method to estimate effects of programs when experimental designs are infeasible is through the use of propensity score analysis (Ravallion, 2001; Rosenbaum, 1984, 1987, 2002; Rosenbaum & Rubin, 1983, 1984, 1985a, 1985b; Rubin, 1980, 2001). Unlike other quasi-experimental designs, which generally rely on a few demographic variables for matching a comparison group to the treatment group, propensity scoring matches treatment and comparison units, incorporates a wide range of variables that have been observed (referred to as covariates) in the matching process. Differences that occur between the two groups, if any, can be estimated from covariates and accounted for through adjustments applied in the form of weights. That is, propensity score matching allows us to address the following question: What would have happened to the Head Start participants if they had participated in Pre-K? Propensity score matching has been shown to be superior to many other approaches, such as multiple regression, because program outcomes can be modeled independently of assignment to treatment across all observed covariates which are included in the matching process.

In our case, we matched children attending Head Start with a group of children attending Pre-K, who could have been eligible to attend Head Start. The multiple covariates used to match children with propensity scoring are converted to a single score using a logit model and the children are matched on the predicted probability or logit that models the propensity of the children to have been enrolled in Head Start (Rosenbaum & Rubin, 1984). Each child in the treatment group (in this case, Head Start) is matched to a child or children in the comparison group (in this case, Pre-K) using a matching method (e.g., within preset caliper width, Mahalanobis metric, or nearest neighbors). Some cases may be excluded from the match if their propensity score is outside the acceptable range of the matching algorithms, while other children could be matched with multiple children. We now turn to a more detailed description of the propensity score method.

Let i index the population under consideration, $i(I = 1, \dots, N)$; Y_{i1} be the value of a standardized test score for child i who attends Head Start (the treatment, $Z_i = 1$); and Y_{i0} be the value of a standardized test score for the same child attending Pre-K (the control, $Z_i = 0$). The treatment effect for individual i is defined as: $\tau_i \mid_{Z=1} = E(Y_{i1} \mid Z_i = 1) - E(Y_{i0} \mid Z_i = 0)$, where the treatment effect is the difference in the expected value of the standardized test result conditioned on the child attending Head Start. However, since the child is not assigned randomly to Head Start or Pre-K, for the child attending Head Start one can only observe $E(Y_{i1} \mid Z_i = 1)$, not $E(Y_{i0} \mid Z_i = 0)$ the test score one would observe had the child been randomly assigned to the alternative treatment, in this case Pre-K.

In a random assignment the expected outcome of the treatment and alternative treatment (or control) are independent, $E(Y_{i0} \mid Z_i = 0) = E(Y_{i0} \mid Z_i = 1) = E(Y_i \mid Z_i = 0)$, mitigating the need to condition on the treatment. Rosenbaum and Rubin (1983) proposed the propensity score, a single dimensional variable that incorporates all available information to form groups that are independent of assignment to treatment based on the observed covariates. One uses the pretreatment covariates to estimate the probability that a child attended Head Start using the propensity score, $p(x)$, such that $p(x) = \Pr\{Z_i = 1 \mid X_i\}$. One note of caution, the bias reduced by the use of the propensity score is limited by the quality and quantity of covariates used to generate the propensity score. Only if the possibility of treatment is random among individuals who have the same propensity score can one say that all bias is eliminated. According to the balancing hypothesis of Rosenbaum and Rubin, they showed that if $p(x)$ is the propensity score, then the treatment assignment is independent of the covariates conditioned on the propensity score $[Z \perp X \mid p(x)]$. Children with the same propensity score have the same distribution of pretreatment covariates independent of treatment assignment. Rosenbaum and Rubin further showed that the assignment to the treatment group is unconfounded when conditioned on the propensity score. Specifically, randomization implies that outcomes are independent of treatment assignment $(Y_{i1}, Y_{i0} \perp Z_i)$, but Rosenbaum and Rubin demonstrated that independence holds if the treatment is conditioned on the propensity score $[Y_{i1}, Y_{i0} \perp Z_i \mid p(x)]$.

Propensity score matching consists of two parts. First, one calculates the propensity score based on the previously discussed probability model using a logit model. The model included all Head Start children active in the study and only those Pre-K children whom we identified as receiving or potentially eligible for government assistance. The result of the logit modeling is displayed in Appendix A. The model explained approximately 40% of the variance in participation in Head Start as opposed to Pre-K. The covariates include characteristics related to the child (e.g., sex, race, age), their family (e.g., parent's education, marital status), their school (e.g., sex, race of class) and their county of residence (e.g., race, income distributions) that could be compiled and reasonably associated with an individual child prior to entry into preschool. The number of covariates was more limited in this study than some others that use propensity score matching because of the sample size.

Second, we estimated the Mahalanobis distance[10] between each child in Pre-K and a child in Head Start, where the Mahalanobis estimate includes the multidimensional distance between two children within strata defined by variables that have been shown in previous research to have been critical factors in school performance.[11] The matching process selects similar cases within each multivariate stratum and picks the closest Pre-K children for each Head Start child. The specified strata variables must be limited, since each additional variable exponentially increases the multivariate dimensions. Given the number of children in the study, we established a limit of five dichotomous variables, which included sex, race (African American status), mother with less than high school education, receiving TANF, and mother less than 20 years of age at the time of their first child's birth (teenage mothers). The propensity score accounts for all observed covariates included in the logit analysis and the specified strata covariates balance the data on the variables found to be influential in prior research.[12] Specifically, we randomly ordered the Pre-K and Head Start children and calculated the distance between the first Head Start child and all Pre-K children. The Head Start child was matched with Pre-K children within the same stratum (that is, having the same specified covariates) and that has propensity score within a certain distance from the Head Start child. The distance is referred to as a caliper. A caliper is in standardized units and the wider one sets the caliper, the more likely that two dissimilar children would be matched. Cochran and Rubin (1973) suggested that researchers should use one-quarter of a standard deviation caliper width. In this study, we matched 106 Head Start participants with 201 Pre-K pupils using the recommended quarter standard deviation caliper width.[13] Eight Head Start children were not matched and these unmatched children were more likely than the 106 children in the matched Head Start subsample: (1) to be White; (2) to be in a class with a greater percentage of boys; (3) to live in county with a relatively smaller percentage of 5-year-olds; (4) to live in county with a higher percentage of married households; and (5) to be less likely to have had a hearing test prior to the start of school.

Results of matching procedure

The propensity score matching decreased significantly the bias that existed initially between

TABLE 2
Reduction in the Bias Between Head Start and Pre-K

Variable	Head Start $N = 106$	Pre-K $N = 201$	% Reduction in bias
Sex: Male = 1			
Initial difference	.55	.45	96.4
After prediction	.55	.55	
Race: African-American = 1			
Initial difference	.62	.50	99.2
After prediction	.62	.62	
Mother's education (% less than high school)			
Initial difference	.74	.70	100.0
After prediction	.74	.74	
Received TANF			
Initial difference	.90	.77	98.6
After prediction	.90	.90	
Married or living with birth father: 1 = Yes			
Initial difference	.40	.54	45.6
After prediction	.40	.32	
Teen mother: 1 = Yes			
Initial difference	.38	.57	87.3
After prediction	.38	.41	

the Head Start and Pre-K samples (see Table 2). The "Initial Difference" results provide demographic information prior to the propensity score matching for Head Start and Pre-K children, while the "After Prediction" results provide post-matching demographics. For example, the initial pool of boys in the Pre-K sample was smaller (45%) than the Head Start sample (55%). Since boys at this age frequently perform worse on standardized tests, this difference may lead one to conclude incorrectly that, on average, children in Pre-K outperform children in Head Start. The reduction in bias due to the propensity score match overweighted the boys participating in Pre-K, to make the overall sample even in the two groups. Other initial differences between the groups were reduced substantially by the propensity matching. Other initial differences included: (1) the Head Start children were far more likely to be African American (62%) than the Pre-K children (50%); (2) the parents of the children in both groups had similar levels of education, with the percentage of parents with less than a high school diploma averaging about 70%; (3) 90% of all Head Start children received TANF, and slightly less than 80% of Pre-K children received TANF; (4) over one-half of the Pre-K children had parents who were married or the birth father was living at home, while that was true for only 40% of the Head Start families; and (5) close to 60% of the Pre-K children came from homes where their mother had her first child when she was a teenager, whereas less than 40% of Head Start children's mothers were a teenage at her first birth. The only meaningful difference that remained after the matching is the percentage of children living in homes where the mother is married or living with the birth father is higher for the Head Start sample. This analysis indicates substantial similarity between the two groups, although clearly not complete equivalence on these covariates.

Rubin (2001) proposed three tests that had been developed to assess equivalence between two groups where regression-based controls were being used, suggested guidelines for appropriate differences, and evaluated a propensity score-matched sample comparison group for a very large data set which he has used to assess the health outcomes of smoking. The three tests compare the differences in the means and variances for propensity scores between the two groups and the residuals for each covariate after regressing them on the propensity score. For this study, the difference in propensity score means for the two subsamples was slightly more than one standard deviation unit, which is greater than the ideal of 0.5 standard deviation unit. The second test required the calculation of the ratio of the variance of the propensity scores for both samples. The ratio was approximately 2.05 (the inverse was 0.49), which was slightly outside the 0.50–2.00 range suggested by Rubin (2001) in which 1.00 or equal variance is the ideal. In the final test, we estimated the variances of residuals from a regression where the propensity score is regressed on each of the variables used to create the propensity score. The average of these residuals was approximately 0.67 where recommended ratio was to fall between 0.50–2.00 with 1.00 being the ideal. While the matched groups did not obtain the ideal, the test statistics fell within or barely outside the recommended ranges. Improvements in the equivalence, given the relatively small sample size for this study, were only achievable by excluding more of the Head Start sample, which we considered a threat to generalizability. Therefore, we proceeded with the analysis using these two groups that we have described.

The third and final test of group equivalence was a test of the mean differences on baseline measures that were taken at the beginning of preschool for both samples. These tests were applied using the probability weights that are described in the next section. The results of the tests are presented in the findings section (Tables 4–7), to make any initial differences in the groups' average skills at the time of entry into preschool apparent when differences in their means at the end of preschool and beginning of kindergarten are presented.

Creating the probability weights

Probability weights are used to adjust scores such that the children in Pre-K who are most similar to the Head Start sample receive greater weights and are more influential in the calculation of the adjusted means. Once the cases are selected, the logit coefficients are converted into probability scores $\left(p = \frac{p}{1+p}\right)$. Probability scores that are less than .05 or greater than .95 are trimmed to .05 or .95, respectively, in order to maximize the weight of any single case at 20. The probability scores of the treatment (Head Start) cases are converted into probability weights $\left(pw = \frac{1}{p}\right)$, while the probability scores of the control cases (Pre-K) are converted into probability weights using the formula, $pw = \frac{p}{1-p}$. Subtracting p from 1 weights the control cases that are most like the treatment cases more, and weights the control cases that are least like the treatment cases less (Foster, 2003).

After selecting the matched cases and generating the propensity score weights, the Head Start and Pre-K test score means were estimated using the weights as adjustments. The test score mean estimates were weighted by the probability weights, which were calculated from the propensity score weights discussed above. The cluster effects of the classroom and the strata effects of the county size were used to adjust the standard errors. The standard errors were estimated using a bootstrap technique (Dehejia & Wahba, 2002). The differences in program quality are estimated using individual level data with standard errors adjusted for the clustering of children within classes in order to use the propensity score weights that were generated at the individual level.[14]

Findings

Program Quality

The first of this study's three measures of program quality was the Early Childhood Environmental Rating Scale-Revised (ECERS-R) scale (1–7), which establishes 5 as the minimum score to be considered good and 3 as the score for minimal quality. Neither the Head Start nor the Pre-K sites serving economically disadvantaged children attained an average score of 5 (see Table 3). On this widely used measure of quality, the two programs seem to be of similar quality, with Pre-K having a statistically insignificant advantage (4.56 versus 4.09).[15] Most of the teachers of economically disadvantaged children in Georgia's Pre-K program have an undergraduate degree (73%). Significantly more Pre-K teachers had at least a bachelor's degree than did Head Start teachers (9%). Finally, almost 28% of the Head Start sites in the study were accredited by the National Association for the Education of Young Children (NAEYC) compared with only 4% of the Pre-K sites. It should be noted that the Pre-K sites located within public schools, which account for almost one-half the sites, are not eligible for NAEYC accreditation. Overall, from a quality standpoint, the Pre-K classes have more highly educated teachers, more Head Start classes have received accreditation from NAEYC, and the quality of their classroom environments appears to be similar.

Direct Assessments

We assessed both groups of children at the start of preschool to establish a baseline; in the spring near the end of their preschool; and again at the start of kindergarten. We have multiple observations for each child on four standardized assessments: the Peabody Picture Vocabulary Test (PPVT), the Woodcock Johnson Test of Achievement using Letter-Word and Applied Problems; and the Oral and Written Language Scales using the Expressive sub-test (OWLS; see Tables 4–7). For sound matching and elision, we assessed the children for the first time at the beginning of kindergarten, because the assessment for 4-year-olds was not available at the time the baseline information was collected (Comprehensive Test of Phonological Processing (Wagner, Torgesen, & Rashotte, 1999; see Table 8). The means for each group were modeled or adjusted using the weights derived from the logit model with the adjustment noted in the methods section. We also adjusted the standard errors for the cluster effects of the

TABLE 3
Preschool Quality With Bootstrap Standard Errors: Propensity Weight-Adjusted Means

Variable	Head Start	Pre-K	Difference	Bootstrap standard error
ECERS-R	4.09	4.56	–0.48	0.31
Teachers with BA	0.09	0.73	–0.64**	0.08
NAEYC accreditation	0.28	0.04	–0.24**	0.07

**$p < .05$. Tests of significance are indicated next to the differences.

TABLE 4

PPVT Scores at Fall and Spring of Preschool Year and at Entry to Kindergarten With Bootstrap Standard Errors: Propensity Weight-Adjusted Means

PPVT	Head Start *N* = 106	Pre-K *N* = 201	Difference	Bootstrap standard error
Entry to preschool	84.01	88.42	4.41	2.70
End of preschool	85.87	92.93	7.06**	2.96
Entry to kindergarten	90.18	93.54	3.36**	1.89

$**p < .05$. Tests of significance are indicated next to the differences.

classroom and stratification effects from the original sampling. In addition to the actual differences in the adjusted means, we present an effect size estimate (d) that allows the comparison of the differences between Head Start and Pre-K across assessments.

PPVT

At the beginning of preschool, children enrolled in Head Start lagged their Pre-K counterparts by 4.41 (p = ns; $d = 0.19$, Table 4). By the end of the preschool year, the Pre-K children were scoring about 7.06 points higher than the Head Start children ($p < .05$; $d = .29$). By the beginning of kindergarten, that difference was reduced to a 3.36 difference in favor of the Pre-K children ($p < .05$; $d = 0.21$).

WJ—Letter-Word

With respect to recognition of letters and words (Woodcock Johnson—Letter Word, Table 5), children enrolled in Pre-K began preschool with an insignificant 0.21 advantage over children enrolled in Head Start (p = ns, $d = 0.01$). By the end of the preschool year, the Pre-K children were testing approximately 4.05 points higher than the Head Start children ($p < .10$; $d = 0.23$). By the beginning of kindergarten, the difference had widened to a significant 4.25 ($p < .05$; $d = 0.32$).

WJ—Applied Problems

The gaps between children enrolled in Head Start and the matched sample of Pre-K children widened significantly on this assessment of general cognition (Woodcock-Johnson—Applied Problems). Like the previous assessments, the difference between the two groups at the beginning of their preschool year was insignificant (p = ns; $d = 0.15$; Table 6). However, by the end of preschool, the difference had increased to a significant 3.66, with the Pre-K children scoring higher than the Head Start children ($p < .05$; $d = 0.27$). Although children enrolled in Head Start did make gains over the summer, the differences between the two groups widened further. By the beginning of kindergarten, the difference between the two groups increased to 4.23 ($p < .05$; $d = 0.33$).

OWLS

Pre-K children began preschool scoring 5.45 point higher in expressive language (OWLS) than the Head Start children ($p < .05$; $d = 0.38$; Table 7). By the end of the preschool year, children enrolled in Pre-K outscored their Head Start counterparts by a statistically significant 6.69 points ($p < .05$; $d = 0.50$). Because of time constraints for the testing, the children were not assessed on the OWLS at the beginning of kindergarten.

TABLE 5

WJ—Letter-Word Scores at Fall and Spring of Preschool Year and at Entry to Kindergarten With Bootstrap Standard Errors: Propensity Weight-Adjusted Means

WJ—Letter-word	Head Start *N* = 106	Pre-K *N* = 201	Difference	Bootstrap standard error
Entry to preschool	97.47	97.68	0.21	3.60
End of preschool	97.85	101.90	4.05*	2.10
Entry to kindergarten	100.02	104.27	4.25**	1.58

$**p < .05$, $*p < .10$. Tests of significance are indicated next to the differences.

TABLE 6

WJ—Applied Problems Scores at Fall and Spring of Preschool Year and at Entry to Kindergarten With Bootstrap Standard Errors: Propensity Weight-Adjusted Means

WJ—Applied problems	Head Start $N = 106$	Pre-K $N = 201$	Difference	Bootstrap standard error
Entry to preschool	89.49	92.44	2.95	2.32
End of preschool	91.26	94.92	3.66**	1.66
Entry to kindergarten	93.33	97.56	4.23**	1.55

**$p < .05$. Tests of significance are indicated next to the differences.

Phonemic awareness

At the beginning of kindergarten, assessments of phonemic awareness were added to the assessment battery. The results reveal modestly to significantly higher scores for children who participated in universal Pre-K (Table 8). The standardized scores for each of the two tests range from 0 to 20. On the Elision Test, which measures blended word recognition, the Pre-K children modestly outperformed the Head Start children (8.3 versus 7.8; $p < .10$, $d = 0.23$). On the sound matching test, the Pre-K children significantly outperformed the Head Start children (9.0 versus 8.2; $p < .05$, $d = 0.42$).

Teachers' Ratings and Other Outcomes Measured at Kindergarten Entry

In addition to the language and cognitive outcomes measured through direct assessments, we measured other outcomes that are important indicators of children's development. In this section, we report adjusted group means on kindergarten teachers' ratings of children's prereading, premath, health status, social skills (ethical behavior, respect for authority, and refusal skills), and overall school readiness. The ratings for academic skills (math, counting, reading, writing, and science) indicated that the Head Start children averaged 4.02 on the scale, while the children from Pre-K averaged 4.66, which approached the "good" labeled (Table 9). The differences ranged from 0.57 (d = 0.22) for counting to 0.74 (d = 0.25) for reading, and all of the differences were significant ($p < .05$). Health and well-being (health, appearance, well rested) results generally indicated that Head Start children were, on average, rated as good, though the Pre-K children consistently received higher overall scores. The differences ranged from an insignificant 0.28 difference for appearance ($d = 0.01$) to a significant 0.57 difference on the health rating ($p < .05$; $d = 0.26$).

The rating for intellectual curiosity and attitudes toward schooling (creativity, curiosity, positive attitude towards schooling) were above average for both groups. However, Pre-K children were rated more highly by their teachers on two of the three ratings. Pre-K children were seen as having a more positive attitude towards schooling (+0.80, $p < .05$, $d = 0.39$) and were more curious (+0.66, $p < .05$, $d = 0.28$). In terms of rating of social skills, children who attended Pre-K were rated as behaving more ethically (+0.48, $p < .05$, $d = 0.25$) and having more appropriate refusal skills (+.65, $p < .05$, $d = 0.43$) There was no significant difference on the respect-for-authority rating.

Kindergarten teachers rated the communication skills of Pre-K children higher, on average, than those of Head Start children. Whereas Head Start

TABLE 7

OWLS Scores at Fall and Spring of Preschool Year and at Entry to Kindergarten With Bootstrap Standard Errors: Propensity Weight-Adjusted Means

OWLS	Head Start $N = 106$	Pre-K $N = 201$	Difference	Bootstrap standard error
Entry to preschool	83.68	88.68	5.45**	1.72
End of preschool	84.93	91.61	6.69**	1.62

**$p < .05$. Tests of significance are indicated next to the differences.

TABLE 8
Phonological Test Scores With Bootstrap Standard Errors: Propensity Weight-Adjusted Means

Entry to kindergarten	Head Start $N = 106$	Pre-K $N = 201$	Difference	Bootstrap standard error
Elision	7.75	8.27	0.53*	0.28
Sound matching	8.18	8.79	0.80**	0.23

$^{**}p < .05$, $^{*}p < .10$. Tests of significance are indicated next to the differences.

children's performance averaged about 4 (the point on the scale labeled "average"), the Pre-K children approached an overall "good" rating. On both ratings, Pre-K children outperformed Head Start children by approximately two-thirds of a point ($p < .05$) on communication skills ($d = 0.35$) and positive expression ($d = 0.35$).

For overall readiness, kindergarten teachers generally rated Pre-K participants as good, significantly higher than the children who attended Head Start. The difference between Pre-K and Head Start children was highest for this overall readiness item (+0.84, $p < .05$, $d = 0.32$). The rating of overall readiness may be an important indicator, since teachers may include aspects of children's behaviors and skills that are not included in other measures.

Four findings across the direct assessments with baseline scores were noteworthy. First, at the beginning of their preschool year, the only significant difference between the two groups was in expressive language skills. The smallest difference on the four baseline measures occurred on the assessment of letter and word recognition, which was a statistically insignificant 0.21 points. On the assessment of expressive language the two groups differed by a statistically significant ($p < .05$) 5.45 points, which is nearly one-third of a standard deviation on the OWLS standardized assessment. Thus, on the third of the three types of tests of preprogram attendance equivalence, the groups were statistically similar on three of four baseline scores, which is important for interpreting the differences in measures of school readiness at the beginning of kindergarten.

Second, on all four tests, the children in both samples started preschool below the national norms for their ages on all four assessments; most

TABLE 9
Teacher Ratings With Bootstrap Standard Errors: Propensity Weight-Adjusted Means

Entry to kindergarten	Head Start $N = 106$	Pre-K $N = 201$	Difference	Bootstrap standard error
Math	4.02	4.60	0.58**	0.29
Counting	4.27	4.83	0.57**	0.31
Reading	4.04	4.78	0.74**	0.36
Writing	3.82	4.46	0.64**	0.30
Science	3.98	4.67	0.70**	0.22
Health	4.94	5.51	0.57**	0.26
Appearance	5.63	5.64	0.28	0.19
Well rested	4.94	5.41	0.47*	0.21
Creativity	4.64	4.99	0.35	0.23
Curiosity	4.04	4.70	0.66**	0.28
Positive attitude	4.39	5.19	0.80**	0.25
Ethical behavior	4.34	4.82	0.48**	0.23
Respect for authority	4.81	5.11	0.30	0.28
Refusal skills	3.86	4.51	0.65**	0.18
Communication skills	3.97	4.70	0.74**	0.25
Positive expression	4.23	4.84	0.61**	0.21
Kindergarten readiness	4.25	5.09	0.84**	0.32

$^{**}p < .05$, $^{*}p < .10$. Tests of significance are indicated next to the differences.

dramatically, both groups were approximately one standard deviation below the norm in terms of their receptive vocabulary skills (PPVT). While the gains for children enrolled in Head Start were modest, the gains for children enrolled in Pre-K averaged about 20% of a standard deviation across the standardized assessments. Among the tests that were administered across all three time periods, the smallest overall gain was for Head Start children on the Letter-Word test, in which the Head Start children moved from a standardized result of 97.5 to 100.02 (+2.5, ns). The largest gain for Head Start children was on the PPVT assessment, where the children moved from a standardized result of 84.01 to 90.18 (+6.17, $p < .01$). However, the smallest gain for Pre-K children was on the PPVT assessment, where the Pre-K children moved from an average standardized score of 88.4 to 93.5 (+5.1 points, $p < .001$). The largest gain for Pre-K children was on the Letter Word assessment, where the children moved from a standardized result of 97.7 to 104.3 (+6.6 points, $p < .001$).

Third, the bootstrap standard errors of the tests decreased at each period for each test. For example, the standard error for the recognition of letters and words was 3.60 at preschool entry, 2.10 at the end of preschool, and 1.58 at the beginning of kindergarten. The only exception was the slight increase in the standard error for the PPVT from the beginning of preschool (2.70) to the end of preschool (2.96). By the beginning of kindergarten, the standard error fell to 1.89. The smaller standard deviation contributed to but was not entirely responsible a statistically significant difference between the groups on at least one measure at the beginning of kindergarten.

Finally, the differences between children attending Pre-K and Head Start widened as the children progressed, though the rate of change may have diminished during the summer between preschool and kindergarten. For example, on the assessment of letter and word recognition, the Pre-K children outscored scored Head Start children by 0.21 points at the beginning of preschool. That difference increased by the end of preschool to 4.05 points. By the beginning of kindergarten, Pre-K children had increased the difference to 4.25. The only exception was on the PPVT test, where the difference between Pre-K and Head Start children diminished during the summer between preschool and kindergarten (7.06 to 3.36). Averaging across the three tests (PPVT, Letter-Word, and Applied Problems) used at the beginning of preschool, the Pre-K children entered kindergarten more than one-quarter of a standard deviation (d = .29) above the children who attended Head Start.

The consistency of the differences is noteworthy between the two groups on the other assessments and on the teachers' ratings, in terms of their direction, statistical significance, and effect sizes. The differences in the assessments of phonological awareness were similar in magnitude to the differences in the other assessments at the beginning of kindergarten, averaging $d = 0.33$. Across all 17 teachers' rating, the children attending Pre-K received higher ratings than Head Start, although three of these were statistically insignificant, including appearance, curiosity, and respect for authority. The effect sizes for the differences in the teachers' rating of the 14 measures that were significant, were generally about one-quarter standard deviation or larger.

Discussion

This study compares the program quality and outcomes of a state prekindergarten program with those of Head Start by taking advantage of the current fragmentation of early childhood education programs. In Georgia, parents have a wide range of choices for their children's preschool. The Georgia Pre-K program provides a developmentally oriented early education program, which is of comparable quality to the programs in 37 other states, for at least 6.5 hours per day over the 180-day school year. The services are frequently monitored, with results for each classroom posted on the Internet, and on-site technical assistance is routinely provided. The federal Head Start Program provides comprehensive services for the children and their families, as well as preschool services. Most of the Head Start programs in Georgia, including all of those in this study, offer preschool services for at least 6 hours a day, 5 days per week. Other options include private preschools and, of course, informal care or staying with family members.

This study's quasi-experimental design, which utilized propensity score matching resulted in two statistically similar groups based on family characteristics and independent analysis of pre-program participation measures of the children's language and cognitive skills. Both groups of children started with similar average scores on three of four standardized assessments, and both made significant gains on age-adjusted standardized scores by the beginning of kindergarten. Children who attended Pre-K began kindergarten better prepared than children who attended Head Start on all six standardized assessments of skills and 11 of 14 ratings by their kindergarten teachers. Since several Head Start agencies in Georgia also provide Pre-K, the differences appear to be related to policy and programmatic differences, not necessarily to the program operator. The differences in developmental status after program participation may be attributable to: (1) policy instrument, (2) program priorities, (3) the program model, (4) the quality of the program implementation, (5) resources available for instruction, (6) peer effects, or (7) program monitoring and oversight. In this study, we cannot attribute the differences to the any specific aspect of the policy or its implementation. It is important to note that Head Start is a more comprehensive program and has objectives for parental involvement and parental self-sufficiency that are not addressed in this study. However, it is possible for a devolved federal program to retain these objectives and mandate certain services and parental outreach activities for the states to implement.

Overall, we conclude from this study that economically disadvantaged children attending a state prekindergarten program were at least as well prepared for school when they entered kindergarten as were the children who attended Head Start. It may be that initial differences explain a part of the differences in children's developmental status after program participation. The study employed a state-of-the-art technique, propensity score matching, to make the observed differences in outcomes independent of the variables influencing Head Start attendance. However, as with any study that does not use and maintain random assignment to treatment, bias is possible. The differences in the assessments at the beginning of preschool provide an independent test of the differences in the two groups. Three out of four indicate that the differences are not statistically significant; but one indicated a significant difference favoring the state prekindergarten program, and on all three of the other assessments, the children attending the state prekindergarten insignificantly outperformed the children attending Head Start. A realistic and important reduction in the potential bias was achieved through propensity score matching, but the ideal of equivalence was not fully achieved.

This study points out the limitations that are likely to occur when using propensity score matching with small samples. Small samples affect not only power to detect post-program differences but also the number of subjects available for matching in the Pre-K sample and the number of variables that could be used to estimate the propensity score and stratify the Head Start sample. Clearly, many large-sample education panel studies offer the potential for fruitfully applying propensity score matching. However, propensity score matching should not be viewed as an easy panacea for the problems in conducting randomized experiments. Larger samples are important in studies using propensity score matching for the reasons cited above, and collecting data on many pretreatment covariates to control for the selection on the observables can be challenging. If future studies are conducted using propensity score matching for these groups, it will be important to consider expanding the sample sizes for both groups in order to take full advantage of the matching technique, expand the number of covariates used to estimate propensity, and, ultimately, improve on the balance between the two groups.

To better understand the implications of the potential devolution of Head Start, a number of additional studies can be suggested. In the immediate future, research using other quasi-experimental techniques, such as regression discontinuity designs (Gormley et al., 2005), should be considered to compare the outcomes associated with Head Start to those of state prekindergarten programs. Using other quasi-experimental methods may reduce or potentially eliminate sources of bias that can occur with any single study or the reuse of the same methods in replication studies. In addition, incorporating standardized health and well-being measures and measures of parental involvement may add to our understanding of differences between state prekindergarten programs and Head Start. Ultimately, any quasi-experimental technique must contend with the possibility of bias and perceptions of bias that can be eliminated only by the use and careful maintenance of random assignment to alternative treatment studies.

The policy debate concerning devolution should incorporate multiple dimensions, including but certainly not limited to children's language, cognitive, and social outcomes included in this study.

Conspicuously absent from this study are cost estimates for the alternatives. Cost, along with feasibility and likely effectiveness, is a common criterion used for policy analyses (Bardach, 2002; Kraft & Furlong, 2004). For this debate, cost information is particularly important because costs of the alternative programs may greatly affect how many children could be served assuming equivalent levels of expenditures. In lieu of accurate cost information, the only fiscal comparisons that are currently available are of per-child expenditures. Using the most recent available estimates, the National Institute for Early Education Research reports that the average per-child-served expenditure for Head Start is $7,089, the average per pupil expenditure for the Georgia Pre-K Program is $3,824, and the median state expenditure for prekindergarten is $3,306 per pupil. Unfortunately, the expenditure data do not include program subsidies from other federal programs, such as Medicaid, or from local contributions, such as local school systems which operate either state prekindergarten programs or Head Start programs. An important direction for future research should be to collect and report accurate data on the full cost of services provided by both Head Start and state prekindergarten programs.

The results from this study, and our observations in the field, have led us to conclude that state prekindergarten programs can deliver early childhood education services that are on par with those delivered by Head Start, and that the developmental status of children served by state prekindergarten programs is at least as high as that of children who have participated in Head Start when they enter kindergarten. These results may be viewed as reducing the perceived risk to the developmental status of the children served by Head Start if devolution were pursued. The developmental status of a matched sample of children served by a state prekindergarten program was higher on most measures. However, this study did not compare health-related screenings or referrals for special services, which are routinely performed by both programs, because the data were not available. Parental engagement in the program and their future advocacy on behalf of their children has been raised as a risk of devolution (Ripple, Gilliam, Chanana, & Zigler, 1999). In addition, parental involvement, which was not evaluated in this study, has benefits for children's educational and social outcomes, (Reynolds, 2000; Reynolds et al., 2001).

Of course, this study was limited to comparing the quality and outcomes of one state prekindergarten program to those of Head Start in one southern state and should be interpreted in that light. We believe that the results are sufficient to encourage additional quasi-experiments and ultimately, randomized experiments in which the coverage, outcomes, quality, and costs of Head Start are compared with state prekindergarten programs. Such experiments could inform future debates concerning the direction of federal policies for improving the social, cognitive, and language skills and well-being of children from economically disadvantaged families.

Notes

[1] Low-income families are defined as family whose earnings total 0–200% of poverty. This is a slightly stricter definition of "at-risk" than that of DECAL. In Georgia, children are eligible for PeachCare (Georgia's SCHIP program), and therefore considered "at-risk" with income levels up to 235% of poverty.

[2] On July 1, 2004, the Office of School Readiness (OSR) received a new name and new responsibilities. OSR is now Bright from the Start: Georgia Department of Early Care and Learning (DECAL). Information and publications gathered from DECAL before July, 2004 are credited to OSR.

[3] For example, payments for students in a classroom with a teacher certified in early childhood education are slightly greater than payments for students who have a lead teacher with lesser credentials. In addition, DECAL funds transportation subsidies ($165 per student per year) for children classified as economically disadvantaged and provides grants to centers serving children living in poverty to help children and their families obtain health and social services and transition to kindergarten.

[4] The Georgia Early Childhood Study includes a cohort of children who attended private preschool not funded by the state or federal government (though the children may individually be subsidized) or who did not attend any formal preschool program (started formal schooling in kindergarten), but these children were not eligible for the analysis reported in this article.

[5] The percentage of Pre-K children classified as potentially eligible for Head Start is roughly the same percentage (52% versus 56%) that was found in a previous study of the Georgia Pre-K program (Henry, Gordon, Henderson, & Ponder, 2003).

[6] On average, there was little difference between when the Head Start and Pre-K children were tested. The widest difference was during the kindergarten year, when we tested the Pre-K children 6 days later on average.

[7] It is important to note that nearly all the classrooms had a 10:1 ratio of student to teachers; therefore, variation was too constrained to make a meaningful comparison of ratios and class size.

[8] The TANF data we received contained multiple formats and coding errors that made it difficult to verify the data. Some TANF personnel entered the data last name first and first name last, while others did the opposite. In some cases, parents provided their own social security number instead of their child's or, if the parents had multiple children, social security numbers were transposed among children. We used data from two different periods to cross-reference the data and account for any changes that TANF personnel may have made along the way. In 70% of the cases, we matched TANF records (or the lack thereof) at the two points. In an additional 8% of the cases, parent surveys confirmed that a child was on TANF as of March 2001. Another 17% of the cases were confirmed by the parents that the child was on TANF at some point on or before December 2003. In the remaining 6% of the cases, the TANF records suggested that the children were on TANF at one point or the other, though the parents did not provide confirmation. In 70% of these cases, the TANF record as of March 2001 was the record indicating that the child was on TANF.

[9] Multiple imputation models provide unbiased estimates of the missing data, even in the face of substantial attrition (Graham, Hofer, & Piccinin, 1994). Most traditional methods assume that data are missing completely at random (MCAR), meaning that missing data are random. A more realistic assumption for missing data is that the data are missing at random (MAR), meaning that the missing data do not depend on the unobserved determinants of the outcome of interest (Little & Rubin, 1987). If this assumption is not valid, then the missing data are described as nonignorable. However, multiple imputation methods have been found to work well with nonignorable missing data (Schafer, 1997). To account for missing data common to survey research, NORM (Schafer, 2000) was used to impute missing values using a multiple imputation algorithm using all available data (see also King, Honaker, Joseph, & Schieve, 2001; Schafer, 1997; Schafer & Graham, 2002).

[10] The Mahalanobis distance is a distance measure based on correlations between the variables and by which different patterns could be identified and analyzed with respect to the base or reference point (Taguchi & Jugulum, 2002).

[11] While there are other methods of matching children, Rosenbaum and Rubin (1985b) suggest that the Mahalanobis match with calipers is preferable.

[12] If one wanted to solely match on the propensity score, then the nearest neighbor match is the preferred method. If the research believes that some covariates are of greater importance, the Mahalanobis method permits the addition of these added factors.

[13] We used the psmatch2 program running in STATA to calculate the propensity scores and matches.

[14] An analysis of adjusted means was preferred to a growth curve analysis for two reasons. First, growth curves require at least three data points, which would have reduced the number of developmental outcomes to three direct assessments and would have left out other skills and the children's overall readiness. Second, the growth curves were nonlinear and would have required at least four data points to conduct a nonlinear analysis.

[15] To account for the clustering of some children in a single class, we weighted the data by the inverse of the number of children in each class for all three variables and multiplied by the probability weights.

References

Advisory Committee on Head Start Research and Evaluation. (1999). *Evaluating Head Start: A recommended framework for studying the impact of the Head Start Program.* Washington, DC: U.S. Department of Health and Human Services.

Bardach, E. (2002). *A practical guide for policy analysis: The eightfold path to more effective problem solving.* New York: Chatham House Publishers.

Barnett, S., Hustedt, J., Robin, K., & Schulman, K. (2004). *The state of preschool: 2004 state preschool yearbook.* New Brunswick, NJ: National Institute of Early Education Research.

Barnett, S. (1991). Benefits of compensatory preschool education. *Journal of Human Resources, 27*(2), 279–312.

Butler, A., & Gish, M. (2003). *Head Start: Background and funding* (No. RL30952). Washington, DC: Congressional Research Service.

Carrow-Woolfolk, E. (1995). *Oral and written language scales.* Circle Pines, MN: American Guidance Service.

Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankya, Series A, 35,* 417–446.

Cochran, W. G. (1965). *Sampling techniques,* 2nd ed. New York: Wiley.

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

Cook, T. D. (2002). Randomized experiments in education: Why are they so rare? *Educational Evaluation and Policy Analysis, 24*(3), 175–199.

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics, 84*(1), 151–161.

Dunn, L. M., Dunn, L. L., & Dunn, D. M. (1997). *Peabody picture vocabulary test—III performance* *record, form A.* Circle Pines, MN: American Guidance Service.

Foster, E. M. (2003). Is more treatment better than less?: An application of propensity score matching. *Medical Care, 41*(10), 1183–1192.

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

Georgia Head Start Collaboration Office. (2003). *Georgia Head Start facts and figures.* Retrieved July 30, 2003, from http://www.osr.state.ga.us/headstart1.html

Georgia Office of Educational Accountability. (2003) Retrieved March 4, 2004, from http://reportcard.gaosa.org/

Georgia Office of School Readiness. (2003). Retrieved March 4, 2004, from http://www.osr.state.ga.us

Gilliam, W. S., & Zigler, E. F. (2001). A critical meta-analysis of all evaluations of state-funded preschool from 1977 to 1998: Implications for policy, service delivery, and program implementation. *Early Childhood Research Quarterly, 15*(4), 441–473.

Gormley, W., & Gayer, T. (2005). Promoting school readiness in Oklahoma: An evaluation of Tulsa's pre-k program. *Journal of Human Resources, 40,* 533–558.

Gormley, W., Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal pre-k on cognitive development. *Developmental Psychology, 41*(6), 872–84.

Graham, J. W., Hofer, S. M., & Piccinin, A. M. (1994). Analysis with missing data in drug prevention research. In L. M. Collins & L. Sietz (Eds.), *Advances in Data Analysis for Prevention Intervention Research* (Vol. NIDA Research Monograph Series (#142)). Washington, DC: National Institute on Drug Abuse.

Grissmer, D., Flanagan, A., Kawata, J., & Williamson, S. (2000). *Improving student achievement: What NAEP scores tell us.* Santa Monica, CA: RAND. MR-924-EDU.

Hart, K., & Schumacher, R. (2004). *Moving forward: Head Start children, families, and programs in 2003.* Center for Law and Social Policy, Washington, DC, Policy Brief 5.

Head Start Bureau. (2004). *Head Start Program Fact Sheet.* Retrieved from http://www.acf.hhs.gov/programs/hsb/research/2004.htm

Heckman, J., Ichimura, H., Smith, J., & Todd, P. (1998). *Characterizing selection bias using experimental data. Econometrica 66.* 1017–1098.

Henry, G. T., & Rickman, D. K. (2006). Effects of peers on early education outcomes. *Economics of Education Review,* in press.

Henry, G. T., Henderson, L. W., Ponder, B. D., Gordon, C. S., Mashburn, A., & Rickman, D. K. (2004). *Report of the findings from the Early Childhood Study: 2001–2002.* Atlanta, GA: Georgia Office of School Readiness and the National Institute for Early Education Research.

Henry, G. T., Gordon, C. S., Henderson, L. W., & Ponder, B. D. (2003). *Georgia Pre-K longitudinal study: Final report 1996–2001 school year.* Atlanta, GA: Georgia Office of School Readiness.

Henry, G. T., Henderson, L. W., Ponder, B. D., Gordon, C. S., Mashburn, A. J., & Rickman, D. K. (2003). *Report of the findings from the Early Childhood Study: 2001–02.* Atlanta, GA: Georgia Office of School Readiness and the National Institute for Early Education Research.

Kagan, S. L., Moore, E., & Bradekamp, S. (1995). *Reconsidering children's early development and learning: Toward common views and vocabulary.* Washington, DC: National Education Goals Panel, Goal 1 Planning Group.

King, G., Honaker, J., Joseph, A., & Schieve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review, 95*(1), 49–69.

Kraft, M. E., & Furlong, S. R. (2004). *Public policy: Politics, analysis, and alternatives.* Washington, DC: Congressional Quarterly Press.

Little, R. J., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Manski, C., & Garfinkle, I. (Eds.) (1992). *Evaluating welfare and training programs.* Cambridge, MA: Harvard University Press.

Mashburn, A. J., & Henry G. T. (2004) Assessing school readiness: Validity and bias in preschool and kindergarten teachers' ratings. Educational Measurement: Issues and Practices, 23(4).

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

NICHD Early Childcare Research Network, & Duncan, G. J. (2003). Modeling the impacts of child care quality on children's preschool cognitive development. *Child Development, 74*(5), 1454–1475.

NIEER Yearbook (2004). http://nieer.org/yearbook/pdf/yearbook.pdf#page=84

O'Brien, R. W., D'Elio, M. A., Vanden-Kiernan, M., Magee, C., Younoszai, T., Keane, M. J., et al. (2002). *Head Start: A descriptive study of Head Start families: FACES technical report I.* Washington, DC: U.S. Department of Health and Human Services.

Office of Planning Research and Evaluation (2003). Building futures: The Head Start impact study interim report, accessed February 6, 2005: http://www.acf.hhs.gov/programs/opre/hs/impact_study/reports/imptstdy_interim/interim_toc.htm

Peisner-Feinberg, E. S., & Burchinal, M. R. (1997). Relations between preschool children's child-care experiences and concurrent development: The cost, quality, and outcomes study. *Merrill-Palmer Quarterly, 43*(3), 451–477.

Quality Counts. (2002). *Building blocks for success.* Retrieved from http://teachermagazine.com/sreports/qc02/templates/article.cfm?slug=17exec.h21.

Ravallion, M. (2001). The mystery of the vanishing benefits: An introduction on impact evaluation. *The World Bank Economic Review, 15*(1), 115–140.

Reynolds, A. J. (2000). *Success in early interventions: The Chicago child-parent centers.* Lincoln, NE: University of Nebraska Press.

Reynolds, A. J., Temple, J. A., Robertson, D. L., & Mann, E. A. (2001). Long term effects of an early childhood intervention on educational achievement and juvenile arrest. *Journal of the American Medical Association, 285,* 2339–2346.

Ripple, C. H., Gilliam, W. S., Chanana, N., & Zigler, E. F. (1999). Will fifty cooks spoil the broth? The debate over entrusting Head Start to the states. *American Psychologist, 54*(5), 327–343.

Rosenbaum, P. R. (1984). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *Journal of the American Statistical Association, 79*(385), 41–48.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association, 82*(398), 387–394.

Rosenbaum, P. R. (2002). Attributing effects to treatment in matched observational studies. *Journal of the American Statistical Association, 97*(457), 183–192.

Rosenbaum, P. R., & Rubin, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society B, 45*(2), 212–218.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*(387), 516–524.

Rosenbaum, P. R., & Rubin, D. B. (1985a). The bias due to incomplete matching. *Biometrics, 41*(March), 103–116.

Rosenbaum, P. R., & Rubin, D. B. (1985b). Constructing a control group using multivariate matched sample methods that incorporate the propensity score. *The American Statistician, 39*(1), 33–38.

Rubin, D. B. (1980). Bias reduction using Mahalanobis-metric matching. *Biometrics, 36*(June), 293–298.

Rubin, D. A. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology, 2,* 169–188.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data* (Vol. 72). London: Chapman & Hall.

Schafer, J. L. (2000). NORM (Version 2.03). University Park, PA: Pennsylvania State University.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177.

Schweinhart, L., & Weikart, D. (1997). The High/Scope preschool curriculum comparison study through age 23. *Early Childhood Research Quarterly, 12,* 117–143.

Shaw, G., and Reinhart, S. (2001). The polls–Trends devolution and confidence in government. *Public Opinion Quarterly 65*(3), 369–389.

Taguchi, G., & Jugulum, R. (2002). *The Mahalanobis-Taguchi strategy: A pattern technology system.* New York: Wiley.

U.S. General Accounting Office. (1997). *Head Start: Research provides little information on impact of current program.* Washington, DC: Report #GAO-HEHS-97-59.

Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1999). *Comprehensive test of phonological processing.* Austin, TX: PRO-ED.

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

Zill, N., Resnick, G., McKey, R. H., Clark, C., Connell, D., Swartz, J., O'Brien, R., & D'Elio, M. A. (1998). Head Start performance measures: Second progress report. U.S. Department of Health and Human Services.

Authors

GARY T. HENRY is a Professor of Public Administration and Urban Studies, Political Science, and Educational Policy Studies, Georgia State University, P.O. Box 4039, Atlanta, GA 30303; gthenry@gsu.edu. His areas of specialization are education policy; school accountability; evaluation of education reform; polling, surveys, and public opinion; and evaluation and policy studies.

CRAIG S. GORDON was a Research Associate, Georgia State University, Atlanta, GA; csgordon@gordonholdings.com. His area of specialization is evaluation methodology.

DANA K. RICKMAN is a Senior Research Associate, Andrew Young School of Policy Studies, Georgia State University, P.O. Box 4039, Atlanta, GA 30303; drickman@gsu.edu. Her areas of specialization are education policy and evaluation and policy studies.

Manuscript received February 14, 2004
Final revision received January 25, 2006
Accepted January 30, 2006

APPENDIX A

Logit Model Predicting Head Start Participation

Logit estimates	Number of observations	=	315
	Likelihood ratio $\chi^2(32)$	=	167.28
	Log likelihood	=	−122.53004
	Pseudo R^2	=	0.4057

Variable	Coefficient	SE	T-Score
Sex (boy = 1)	2.32	1.89	1.23
Black	0.00	0.91	0.00
Black* % of county Black	0.05	0.03	1.70
Other race* % of county other	0.05	0.04	1.29
Speech defect	−1.70	0.63	2.71
Age	35.62	21.74	1.64
Age^2	−4.02	2.44	1.65
Mother education < high school	0.27	0.45	0.60
Father education < high school	−0.59	0.41	1.43
Eye exam before prekindergarten	0.99	0.83	1.18
Ear exam before prekindergarten	−1.82	0.93	1.97
Parent chose program because of perceived quality	−0.59	0.58	1.00
Parent chose program because of social interaction	−1.55	0.72	2.17
Parent chose program because of location	0.13	0.48	0.27
% of class boys	0.14	0.03	4.90
% of class Black	0.06	0.02	3.60
% of class other race	0.07	0.02	4.56
Boy* % of class boys	−0.04	0.03	1.02
Black* % of class Black	−0.02	0.02	1.21
% of county under 5	−0.42	0.25	1.70
% of county Black	−0.10	0.03	3.02
% of county Asian	−0.23	0.12	1.91
% of county Hispanic	−0.12	0.06	2.15
% of county married	0.01	0.03	0.56
% of county grandparents head household	0.31	0.24	1.26
% of county with income < $25,000	0.21	0.08	2.67
% of county with income $25,000–$35,000	0.18	0.10	1.92
% of county with income $35,000–$50,000	−0.26	0.09	2.86
% of county living in the same house for 5 years	−0.14	0.03	4.41
% of females working out of the home in the county	0.12	0.10	1.21
% of county commuting out of the county	0.09	0.04	2.36
% of county without phone service	0.13	0.08	1.75
Constant	−86.44	48.35	1.79

APPENDIX B
Comparison of Means Before and After Propensity Score Matching

Test & Period	Head Start (no matching)	Head Start (matching)	Pre-K (no matching)	Pre-K (matching)
PPVT (fall Pre-K)	82.35	84.01	90.10	88.42
PPVT (spring Pre-K)	85.24	85.87	93.75	92.93
PPVT (fall K)	93.82	90.18	93.78	93.54
Letter-word (fall Pre-K)	94.84	97.47	100.70	97.68
Letter-word (spring Pre-K)	96.17	97.85	102.31	101.90
Letter-word (fall K)	98.98	100.02	104.43	104.27
Applied problems (fall Pre-K)	88.94	89.49	94.43	92.44
Applied problems (spring Pre-K)	90.88	91.26	96.30	94.92
Applied problems (fall K)	93.04	93.33	98.29	97.56
OWLS (fall Pre-K)	83.31	83.68	89.43	88.68
OWLS (spring Pre-K)	84.53	84.93	91.96	91.61
Elision (fall K)	7.86	7.75	8.42	8.27
Sound matching (fall K)	8.20	8.18	9.04	8.79