Published in final edited form as:
J Spec Educ. 2010 February 1; 43(4): 236–254. doi:10.1177/0022466908323007.

A Propensity Score Matching Analysis of the Effects of Special Education Services

Paul L. Morgan, **Michelle Frisco**, **George Farkas**, and **Jacob Hibel**
The Population Research Institute & The Pennsylvania State University

Abstract

We sought to quantify the effectiveness of special education services as naturally delivered in U.S. schools. Specifically, we examined whether children receiving special education services displayed (a) greater reading or mathematics skills, (b) more frequent learning-related behaviors, or (c) less frequent externalizing or internalizing problem behaviors than closely matched peers not receiving such services. To do so, we used propensity score matching techniques to analyze data from the Early Childhood Longitudinal—Study Kindergarten Cohort, 1998–1999, a large scale, nationally representative sample of U.S. schoolchildren. Collectively, results indicate that receipt of special education services has either a negative or statistically non-significant impact on children's learning or behavior. However, special education services do yield a small, positive effect on children's learning-related behaviors.

Special education provides children with disabilities with specialized services designed to "prepare them for further education, employment, and independent living" (Individuals with Disabilities Improvement Education Act, 2004). Practitioners are responsible for providing specific services, instructional strategies or routines, and resources that mitigate the impact of the disability on a child's learning or behavior. Doing so should allow the child to better access his or her school's general curriculum, or develop the skills and competencies necessary to participate in that curriculum (Bateman & Linden, 2006). Helping the child to benefit from the school's curriculum should in turn increase his or her subsequent educational and societal opportunities (U.S. Department of Education's Office of Special Education Programs, 2006). Over 6 million children and youth receive special education services in the U.S. (U.S. Department of Education, 2005).

These services are costly to provide. For example, the federal government spent about $50 billion on special education services in 1999–2000 (President's Commission on Excellence in Special Education, 2002). In contrast, the government spent $27.3 billion and $1 billion, respectively, to fund regular education services and other additional special programs (e.g., Title 1). In per-pupil expenditures, the federal government spent about $8,080 on special education services and $4,394 on regular education services. Thus, the U.S. Department of Education (2002) estimated that "the nation spends 90% more on a special education student than on a regular education student (p. I-24)." These estimates do not include additional funds allocated by states or localities. These are also substantial. For instance, in 2008, the state of Kansas expects to pay $404,439,603 to provide students with special education services, after subtracting out the costs of providing regular education services and accounting for funds received in federal aid and Medicaid reimbursements (Kansas Legislative Research Department, 2004).

Correspondence concerning this article should be addressed to Paul L. Morgan, Department of Educational Psychology, School Psychology and Special Education, 211 CEDAR Building, University Park, Pennsylvania, 16802. (814) 863-2285 (Office); (814) 863-1002 (Fax). paulmorgan@psu.ed.

Are Special Education Services Effective?

There is much indirect evidence that, despite their expense, special education services may not be functioning to increase children's educational or societal opportunities. For example, 66% and 68% of eighth grade youth with disabilities scored below the basic level on the 2005 National Assessment of Educational Progress (NAEP) reading and mathematics measures, respectively (U.S. Department of Education, 2007). In contrast, only 24% and 27% of youth without disabilities scored below this level on the two measures. Those youth with disabilities about to enter the nation's job market and postsecondary schools performed even worse. Seventy-two percent and 83% of twelfth graders with disabilities scored below the basic level on the NAEP's reading and mathematics measures. Children and youth with disabilities also continue to lag far behind their non-disabled peers on varying measures of societal attainment (Phelps & Hanley-Maxwell, 1997). For example, those with disabilities are more likely to drop out of school, be delinquent, be unemployed, earn less, and be unsatisfied with their adult lives than their non-disabled peers (Blackorby & Wagner, 1996). Both legal judgments (e.g., Zachary Deal v. Hamilton County Department of Education, 2001, 2004) and classroom observations (e.g., Levy & Vaughn, 2002; Magiera & Zigmond, 2005; Shores, Jack, Gunter, Ellis, DeBriere, & Wehby, 1993) indicate that children with disabilities do not always receive services that can be expected to mitigate the effects of their disabilities. Indeed, children placed in special education classrooms sometimes score lower on measures of reading, writing, and mathematics skills at the end than at the beginning of the school year (Lane, Wehby, Little, & Cooley, 2005). Such findings have led many researchers to attempt to identify ways to increase the effectiveness (e.g., by seeking to close the "research-to-practice gap") of special education services (Abbott, Walton, Tapia, & Greenwood, 1999; Carnine, 1997; Deshler, 2005; Gersten, Vaughn, Deshler, & Schiller, 1997; Snell, 2003).

Yet the effectiveness of special education services remains to be established. This is because prior studies have not used design features (e.g., random assignment, matching) that help control for selection bias. Selection bias is particularly likely to occur in programs intended for special populations. Instead, most studies have relied on simple contrasts between children who received special education services and those who did not (e.g., Blackorby & Wagner, 1996; NAEP, 2005). For example, the NAEP's analyses do not statistically control for variation in children's background characteristics, such as between-group differences in children's initial (i.e., at school entry) level of proficiency in reading, mathematics, or behavior, or in a range of socio-demographic factors (e.g., gender, socioeconomic status), before contrasting the learning and behavior of those who received and who did not receive special education services. Reliance on such simple contrasts is problematic because they likely yield biased estimates of special education's effects. That is, these types of contrasts confound differing characteristics of the groups with differences in each group's outcomes (Shadish, Cook, & Campbell, 2002).

Properly estimating the effectiveness of special education services as naturally delivered in U. S. schools necessitates contrasting outcomes for two groups of children. These are (a) children receiving special education and (b) an *equivalent* group of children not receiving such services. A randomized control trial would best approximate this contrast, in which children were randomly assigned to receive or not receive special education services (Shadish et al., 2002). Use of random assignment should result in children in the treatment and control groups being equivalent on both observed and unobserved background characteristics, thereby controlling for any selection bias. Doing so should yield a relatively unbiased estimate of special education's effects on children's learning and behavior. The study's contrasts, if based on a large-scale nationally representative sample of U.S. schoolchildren, should also allow for generalizations about special education's effectiveness as a federally mandated program of compensatory services. For example, use of such a representative sample should help account for observed large state-level variation in special education placement rates (Donovan & Cross, 2002).

Using Propensity Score Matching to Evaluate Special Education's Effectiveness

Randomly assigning children to receive or not receive special education services is not possible because children meeting eligibility criteria are legally entitled to these services. An alternative method of reducing selection bias is to use propensity score matching techniques. These techniques allow for quasi-experimental contrasts between children in naturally occurring "treatment" and "control" groups, but who display similar likelihoods of experiencing the treatment based on their observed characteristics. Here, we used propensity score matching to contrast the learning and behavior of children who did and who did not receive special education services, but who, as indicated by measures of their health and well-being, family experiences, child-care experience prior to kindergarten entry, socioeconomic background, prior learning and behavior, and quality of school experiences, displayed similar propensities to receive these services.

Proper use of propensity score matching should allow for rigorously derived and relatively unbiased estimates of special education's effects on children's learning and behavior (Dehejia & Wahba, 1998; Hong & Raudenbush, 2005; Shadish et al., 2002). Results obtained from quasi-experiments using propensity score matching methods can closely approximate those obtained from randomized control trials (Becker & Ichino, 2002). For example, Luellen, Shadish, and Clark (2005) contrasted findings from two quasi-experiments using propensity score matching to those obtained from two true experiments. Use of propensity score matching reduced selection bias by 73–90%. The mean differences obtained from the propensity score analyses and those obtained using randomization differed by only .09 and .20 of a point on each study's particular outcome measure. Because of its ability to greatly reduce selection bias, propensity score matching is increasingly being utilized in the fields of policy evaluation (Harknett, 2006; Jones, D'Agostino, Gondolf, & Heckert, 2004), medicine and epidemiology (Stone et al., 1995), economics (Czajka, Hirabayashi, Little, & Rubin, 1992; Lechner, 2002), psychology (Green & Ensminger, 2006) and education (Rosenbaum, 1986; Schneider et al., 2007).

Study's Purpose

We sought to quantify the effects of naturally delivered special education services on children's learning and behavior. Specifically, we used data from the Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), a large-scale, longitudinal, and nationally representative sample of U.S. schoolchildren, to examine whether children receiving special education services in the spring of 2002 (when they were about 8–9 years old) displayed (a) greater reading or mathematics skills, (b) more frequent learning-related behaviors, or (c) less frequent externalizing or internalizing problem behaviors in the spring of 2004 (when they were about 10–11 years old) than peers with similar characteristics who did not receive such services. We used three propensity score matching techniques to estimate the impact of special education services on five measures of children's learning and behavior. We also contrasted these results to estimates obtained using simple tests of mean differences and ordinary least squares (OLS) regression. Our use of a large and longitudinal sample of U.S. schoolchildren, multiple measures of children's learning and behavior, and multiple statistical strategies, should help inform policy-makers, researchers, and practitioners on the magnitude of special education's effects on the learning and behavior of children with disabilities.

Method

Study's Database

The ECLS-K is maintained by the U.S. Department of Education's National Center for Education Statistics (NCES). The ECLS-K is the first large scale, nationally representative sample of children as they age through the elementary school years. The sample was selected to be representative of all U.S. schoolchildren entering kindergarten in the fall of 1998. Children were recruited from both public and private kindergartens offering full- or half-day classes. Data from the sampled children were collected in the fall of 1998, the spring of 1999, the fall of 1999 (with only data from a random sub-sample collected at this time point), and again in the springs of 2000, 2002, and 2004.

Study's Analytical Sample

Our study used two samples of children. We estimated the propensity to receive special education services using data from the first sample. We used data from the second sample to estimate the effects of special education services on children's learning and behavior. To be eligible for the first sample, children had to participate in the ECLS-K in the springs of 2002 and 2004 and have data on whether they were receiving special education services in the spring of 2002 (N=11,479). We further constrained this sample to those children having complete data on all those covariates used to predict their propensity to receive special education services (N=8,020).

Those children eligible for the second sample had complete data on each measure of learning and behavior. Thus, we excluded those children without complete 2002 and 2004 Reading and Mathematics Test scores or ratings of their learning-related, externalizing, or internalizing problem behaviors. This constraint resulted in a final second sample of 6,318 children, 363 of who received special education services and 5,955 of who did not. Most (i.e., 87%) of the children receiving special education services had been identified as having learning disabilities or speech or language impairments. Supplementary analyses indicated that our sample is slightly more advantaged than the full ECLS-K sample. The analytic sample includes a greater proportion of children who are non-Hispanic White, living in households with two biological parents, and whose families had a higher socioeconomic status. Only 7% of our analytical sample's schools were attended by more than one child receiving special education services.[1]

We did not constrain the sample used to construct our propensity score to the smaller sample used to estimate mean differences between treatment and control cases. This is because our goal in using propensity score matching was to estimate children's true propensities, or "known propensities" to receive a treatment. Such estimates will typically have more precision when based on more data (Rosenbaum & Rubin, 1983). Using a larger sample reduces standard errors and in turn leads to a more precise estimate of the propensity score (Frisco, Muller, & Frank, 2007).

Study's Measures

We predicted the effect of special education on two measures of children's learning and three measures of their behavior. Specifically, we examined whether children who were receiving special education services in the spring of 2002 displayed (a) greater reading or

[1] Tables reporting on (a) children's disability identification, (b) descriptive statistics contrasting the study's analytical sample and ECLS-K full sample, (c) the analytical sample's special education placement per school, (d) the propensity score model's covariates, and (e) the relation of the covariates (using logistic regression coefficients) to children's special education placement are available at the study's first author's departmental webpage (http://espse.educ.psu.edu/faculty_web_page.php?id=31).

mathematics skills, (b) more frequent learning-related behaviors, or (c) less frequent externalizing or internalizing problem behaviors in the spring of 2004 than closely matched peers who did not receive such services. Using multiple measures of a program's effects better identifies its potential impact (Gersten, Baker, & Lloyd, 2000). Each of the outcomes is a key contributor to a child's later educational and societal opportunities (e.g., Diperna, Lei, & Reid, 2007; Fergusson & Woodward, 2002; McClelland, Acock, & Morrison, 2006; Rivera-Batiz, 1992; Schaefer, Petras, Ialongo, Poduska, & Kellam, 2003; Schneider, 2001). Below, we detail our measures of children's reading skills, mathematics skills, learning-related behaviors, and externalizing and internalizing problem behaviors.

The Reading Test—The ECLS-K Reading Test seeks to measure children's basic skills (e.g., print familiarity, letter recognition, decoding, sight word recognition), receptive vocabulary, and reading comprehension skills (i.e., making interpretations, using background knowledge). The Reading Test was created through a multi-stage panel review. Some items were borrowed or adapted from published tests (e.g., the Peabody Picture Vocabulary Test—Revised, the Woodcock Johnson Tests of Achievement—Revised). The Educational Testing Service, elementary school curriculum specialists, and practicing teachers supplied other items. Each item was field-tested. Items were included in the Test if they displayed (a) acceptable item-level statistics, (b) good fit with maximum likelihood item response theory (IRT) parameters, and (c) no differential item functioning across gender or race (NCES, 2005). NCES-trained field staff individually administered the Reading Test using an un-timed format. NCES uses a routing procedure (i.e., a child is given a different battery of test items depending on the accuracy of his or her initial responses) and IRT methods to derive scale scores that are then comparable across grade levels. NCES considers reliabilities of the Reading Test's IRT theta scores (i.e., estimates of a child's ability) to be the most appropriate internal consistency estimates. These reliabilities were . 91, .93, .96, .93, and .94 for the fall and spring of kindergarten and the springs of first, third, and fifth grade, respectively (NCES, 2005). First graders' Reading Test scores correlated .85 or above with the Kaufman Test of Educational Achievement reading test (NCES, 2002); third graders' scores correlated .83 with the Woodcock-McGrew-Werder Mini-Battery of Achievement (NCES, 2005).

The Mathematics Test—The Mathematics Test seeks to measure a range of mathematics skills (e.g., identify numbers and shapes, sequence, add or subtract or multiply or divide, use rates and measurements, use fractions, calculate area and volume). NCES used a multi-step panel review process to develop the ECLS-K's Mathematics Test (NCES, 2005). This test was based on the NAEP's specifications. A wide range of kindergarten, first grade, third grade, and fifth grade-level mathematics test bank items were used. NCES also used IRT-methods to create adaptive tests that were administered one-to-one to each child, in an un-timed format. Thus, children were given a test whose coverage varied according to their grade and skill level. Like the Reading Test, the Mathematics Test has strong psychometric properties. For example, reliabilities of the theta scores ranged from .89 to .94 between the fall of kindergarten and the spring of fifth grade (NCES, 2005).

The Teacher Social Skills Rating Scale—The ECLS-K modified the Social Skills Rating System (SSRS; Grehsam & Elliott, 1990) to measure's children's behavior. The original psychometric data of the Social Skills Rating System were based on 4,170 K-12 students (Gresham & Elliott, 1990). Seventeen percent of these students attended special education classes. The test–retest correlation over 4 weeks was .85 (Gresham & Elliott). Correlational and factor analyses support the measures' construct validity (Feng & Cartledge, 1996; Furlong & Karno, 1995). NCES's subsequently modified the SSRS. These modifications included (a) the addition of items measuring the child's frequency of positive affect, behavior, and approaches to learning, (b) expanding the response format from a three point to a four point scale and including a "not observed" response, and (c) re-wording some items to reduce cultural bias (NCES, 2005). Meisels, Atkins-Burnett, and Nicholson (1996) provide additional details on the modifications to the SSRS.

The ECLS-K Teacher SSRS includes three subscales: (a) Approaches to Learning; (b) Externalizing Problem Behaviors; and (c) Internalizing Problem Behaviors (NCES, 2004). Teachers use a frequency scale to rate how often the child displays a particular social skill or behavior (i.e., 1 = student never exhibits this behavior; 4 = student exhibits this behavior most of the time). Items used for the Approaches to Learning subscale measure behaviors impacting how well a child manages his or her behavior while completing learning-related tasks (e.g., remaining attentive, persisting at task, being flexible and organized). The Externalizing Problem Behaviors subscale's items measure acting out behaviors (e.g., arguing, fighting, showing anger, acting impulsively, disturbing the classroom). The Internalizing Problem Behavior subscale's items ask teachers about whether the child appears anxious, lonely, sad, or has low self-esteem. NCES (2005) reports that the split-half reliabilities for the five scales for first grade and third grade children were, respectively, .89 and .91 for Approaches to Learning, .86 and .89 for Externalizing Problem Behaviors, and . 77 and .76 for Internalizing Problem Behaviors. These reliabilities for fifth grade children were .91 for Approaches to Learning, .89 for Externalizing Problem Behaviors, and .77 for Internalizing Problem Behaviors. Exploratory and confirmatory factor analyses confirmed the full scale's structure (NCES, 2005).

Participation in Special Education

We operationalized receipt of special education services as whether a child was being provided with special education services when ECLS-K field staff collected these data in the spring of 2002. NCES examined school administrative records to determine whether or not a child was receiving special education. School administrative records are a frequently used indicator of a child's disability status (e.g., Hollomon, Dobbins, & Scott, 1998; Hosp & Reschly, 2002). This measure of participation in special education included children just beginning to receive special education services in the spring of 2002, as well as children already receiving such services.

We conducted "intent-to-treat" analyses (Lachin, 2000), in which we considered all those children so identified by school districts as receiving special education services, regardless of the children's adherence to the eligibility criteria, what types of services they actually received, and whether they subsequently moved out of special education. Intent-to-treat analyses help account for selection bias and so should result in less biased estimates of a program's effectiveness (Petkova & Teresi, 2002).

Descriptive analyses for the 2002 ECLS-K's full sample of children participating in special education indicated that the mean number of hours per week of special education services received by children was 6.78 (*SD*=9.48). The most commonly used practices, as a percentage of special education teachers who reported using these practices, were small group instruction (95%), direct instruction (86%), and one-to-one instruction (76%). The mean number of hours per week of special education services received by the 2004 full sample was 7.6 (*SD*=11.32). The most commonly used practices by special education teachers remained small group instruction (97%), direct instruction (88%), and one-to-one instruction (78%).

Predictors of Special Education Placement

We used 35 covariates and 4 interaction terms to model a child's propensity to receive special education services. We used both theory (e.g., Kavale, 1988; Stanovich, 1986) and prior empirical research (e.g., Delgado & Scott, 2006; Donovan & Cross, 2002; Mann, McCartney, & Park, 2007) to identify background characteristics that increase a child's risk of being identified as disabled. For example, we used gender as a covariate because boys are two to three times more likely to be identified as learning disabled than girls, regardless of whether a regression-, discrepancy, or low-achievement identification method is used (Katusic, Colligan, Barbaresi, Schaid, & Jacobsen, 2001). We used the child's race, both family- and school-level socioeconomic status, and the parents' marital status as covariates because they are known to predict school-based disability identification (e.g., Hosp & Reschley, 2004; Mann et al., 2007; U. S. Department of Education, 2007). Prior studies also indicate that a child's reading, mathematics, and learning-related behaviors at school entry strongly predict his or her subsequent placement into special education (e.g., Hibel, Farkas, & Morgan, 2008; Mann et al.; Merrell & Shinn, 1990). These factors are also included here. To maximally reduce the potential for selection bias, many covariates should be included in the model predicting propensity to receive treatment (e.g., Shadish et al., 2002), even those that only weakly predict the treatment (Rubin, 1997). Thus, we also included many additional covariates, such as whether the child participated in Head Start, the number of the child's siblings, his or her birth weight, his or her mother's age when she first gave birth, and whether the child had received non-parental child care prior to kindergarten. These factors can also function as predictors of special education placement (e.g., Andrews, Goldberg, Wellen, Pittman, & Struening, 1995; Mannerkoski, Aberg, Autti, Hoikkala, Sarna, & Heiskala, 2007).

Analytical Strategies

Simple mean differences—We used *t*-tests to determine whether mean differences in the learning and behavior of children receiving or not receiving special education services were statistically significant. We used a conservative *p* value of .01 for these analyses. As noted above, this type of analysis is confounded by selection bias. Therefore, we considered the results obtained from *t*-tests of these mean differences as a "benchmark" for the results obtained from analyses using propensity score matching.

OLS regression—We used OLS regression to estimate the average treatment effect of special education services. We used the following OLS regression equation:

$$\mathrm{y}=\alpha+\beta_{1\mathrm{i}}\mathrm{x}_{1\mathrm{i}}+\beta_{2\mathrm{i}}\mathrm{x}_{2\mathrm{i}\ldots\ldots}+e \quad (1)$$

where y = the outcome of interest, α = the constant term, each x_i represents an independent variable associated with outcome y, each β_i indicates the parameter estimate for each x_i, and e represents an error term. The covariates used as control variables in OLS regression models are the same covariates used in models estimating children's propensity to receive special education services (described below).

OLS regression analyses may not always adequately control for selection bias (Heckman, Ichimura, & Todd, 1997; Winship & Mare, 1992). For example, OLS regression's assumption of a linear or nonlinear functional form between an outcome and the covariates may not hold, especially if the covariate distributions differ substantially between the treatment and control groups. We therefore considered results obtained from the OLS regression models as an additional benchmark against which to compare results obtained through use of propensity score matching techniques.

Propensity score matching—We used propensity score matching to allow for contrasts of the learning and behavior of children who did and did not receive special education services, but who, as indicated on the basis of a wide range of observed background characteristics, had the same or nearly the same probability of receiving such services. Propensity score matching yields an average treatment effect for the treated (ATT), which is considered a better indicator of whether to continue policy or programs that target a specific group of interest than the population-wide average treatment effect obtained using OLS regression methods (Heckman, 1996).

Propensity score matching is a two-stage process. Stage 1 involves the use of a logistic or probit regression model to calculate all respondents' propensity for experiencing a treatment of interest, in this case receiving special education services. The propensity score is defined as follows (Rosenbaum & Rubin, 1983):

$$p(T) \equiv Pr\{T=1|S\}=E\{T|S\} \quad (2)$$

where, here*, p(T)* is the propensity to be placed into special education; *T* indicates that a child did or did not receive special education; and *S* is a vector of covariates influencing whether the child did or did not receive special education. We utilized a logit rather than a probit model to predict children's propensity to receive special education services (Goodman, 1978; Rosenbaum & Rubin). The 35 covariates used here to predict children's propensity to receive special education services represent their health and well-being, family experiences, child-care experience prior to kindergarten entry, socioeconomic background, prior learning and behavior, and quality of school experiences. Four interaction terms are also included in the model. We measured all variables except those representing the child's reading or mathematics skills or school transitions using data from the ECLS-K's kindergarten surveys, administered when the children were about 5–6 years of age. Using variables measured at school entry helped avoid the endogeneity problems associated with controlling variables that might themselves have been affected by special education services the student subsequently received, or did not receive, between 1998 and 2002. We estimated a child's initial level of reading or mathematics skills by averaging his or her kindergarten and first grade test scores on the ECLS-K's Reading and the Mathematics Tests. We used the STATA "pscore" command to generate propensity scores (see Becker & Ichino, 2002, pp. 12–14 for an example of this code and output produced by the "pscore" command). This procedure automatically tests for balance between the treatment and control groups on covariates used to predict the propensity score

Table 1 shows the grouping of children into strata based on their predicted propensity scores of special education placement. For the two groups of children in stratum 1 (i.e., the stratum of children with the lowest propensity to receive special education), the propensity to be placed into special education was both .01 for those who were and were not so placed. The propensities for the two groups of children in stratum 8, which consisted of those with the highest estimated propensity to receive special education, were .86 and .84, respectively. Each of the covariates and interaction terms except for the dummy variable representing kindergarten school sector balanced within each stratum. Thus, their mean or modal values were not statistically different across matched treatment and control children. A child's public or private school sector balanced across seven of the eight propensity score strata; it did not balance in the stratum containing cases with the lowest propensity to be placed in special education. Our model and covariate balance is comparable to that reported in other studies using propensity score matching (e.g., Hong & Raudenbush, 2005). Overall, results from Stage 1 of the propensity score matching indicated that our model matched treatment and control children well, thereby reducing the potential for selection bias.[2]

In Stage 2, we used the estimated propensity scores obtained in State 1 to match children who were and who were not placed in special education. We used three matching methods. Doing so is a form of sensitivity analysis that helps guard against hidden bias (Luellen et al., 2005). Each technique uses a different function. However, each yields an ATT attributable to the receipt of special education services. Stratification matching utilizes all treatment and control cases. Using the STATA 9.0 “atts” command, the full range of sample members’ propensity scores is divided into propensity score strata, or blocks, each of which includes treatment and control cases with the same or nearly the same propensities for receiving the treatment. The number of appropriate strata depends upon the number necessary to produce a balanced propensity score. Within each of these strata, the ATT is calculated, and then the ATT’s across strata are averaged to produce a final ATT.

Nearest neighbor matching as conducted using the STATA 9.0 “attnd” command utilizes estimated propensity scores to pair cases in the control and treatment group based on their likelihood of experiencing a treatment. Control cases not matched to a treatment case are excluded from analyses. We used matching with replacement to identify neighbor cases. Matching using replacement allows control cases to be matched to more than one case in the treatment group if that control case is a better match to multiple treatment cases than alternative control cases. Matching with replacement helps minimize bias in the calculation of the ATT (Frisco et al., 2007).

Kernel matching uses the calculated propensity score to match individual cases in the treatment group to a weighted mean of control cases. Control cases receive weights based on the distance between their propensity score and the propensity score of the treatment case to which they are being matched. All control cases can potentially contribute to the weighted mean composite of the control cases, which improves estimation power and efficiency (Frisco et. al. 2007). This is especially important when there are many potential matches for each treatment subject, as was the case with our sample. However, kernel matching results in control observations contributing to more than one match. The matches are not independent, thereby violating an assumption of ordinary parametric techniques for calculating standard errors. Therefore, we bootstrapped using 1000 repetitions to obtain standard errors. We then used the bootstrapped standard errors to make statistical inferences. We used a relatively small bandwidth of .01 when matching cases. Pagan and Ullah’s (1999) exploratory analyses indicated that this leads to a less biased estimate of the ATT.

Results

Below we present results from four types of analyses. The first set of results, displayed in Table 2, are estimates of mean differences in the learning and behavior between the sub-samples of ECLS-K children who received (*n* = 363) and who did not receive (*n* = 5,955) special education services. These children have not been matched on their propensity to receive special education services. The second set of results, reported in the second and third

[2] We used multiple methods to produce our estimates. Each set of analyses utilized data from all 6,318 cases except those conducted using nearest neighbor matching. These were constrained to the 363 children who received special education services and the 274 children who did not receive special education services but who had been “matched” based on observable covariates predicting special education placement. We conducted extensive sensitivity analyses when estimating our final propensity score model. We also evaluated how inclusion and exclusion of different variables, including interaction terms and polynomial terms, influenced our ability to achieve balance, as well as whether use of data trimming (excluding potential outliers) influenced our findings. We conducted sensitivity analyses to investigate whether different bandwidth and bootstrapping specifications affected the robustness of the ATT estimates produced using kernel matching. They did not. In addition, we estimated a second propensity score model that further constrained our sample to those children with complete data on (a) individually-administered measures of their reading and mathematics skills, and (b) teacher-reports on their learning-related behaviors, externalizing problem behaviors, and internalizing problem behaviors (*N*=6,318). This second model was estimated from a smaller sample size and fewer variables, but did balance across all covariates. Results from this second model were consistent with our reported results.

columns of Tables 3 and 5, are obtained using OLS regression methods. The third column in both tables displays results from OLS regression using the same covariates used to estimate a child's propensity to receive special education services. That is, Tables 3 and 5's third column displays differences in the learning and behavior of children who received and did not receive special education, after statistically controlling for the same covariates used to predict children's propensity to these services. The third set of results, shown in the last three columns of Table 3 and 5, are estimates obtained from three different propensity score matching techniques. These results indicate differences in the learning and behavior between children who received and who did not receive special education services, but who have been matched on their propensity to receive such services. The fourth set of analyses, displayed in Tables 4 and 6, disaggregates these differences by estimating special education's effects within each of the eight propensity score strata. Each stratum includes children who received and who did not receive special education but whose estimated propensities to receive such services were the same or nearly the same. For each measure of children's learning or behavior, we report differences between special education and non-special education children in both their (a) spring 2004 scores, and (b) gain scores between the springs of 2002 and 2004. For example, we report whether the children who did or did not participate in special education displayed systematic differences in their reading and mathematics IRT scores in spring 2004, as well as differences in their relative gains in their reading and mathematics IRT scores between the springs of 2002 and 2004. We examined special education's effects on the children's teacher-rated behavior in the same way. We used gain scores to help ensure that any observed differences in spring 2004 between special education and non-special education children who were alike on a large number of background characteristics were not statistical artifacts of initial learning or behavior differences in spring 2002. Gain scores can correctly be used in this way when estimating a treatment effect following implementation of a program or intervention (Allison, 1990; Maris, 1998; Williams & Zimmerman, 1996).

Do Special Education Services Improve Children's Learning of Reading or Mathematics?

Effects on Reading Skills—Table 2 displays simple mean differences between those receiving and not-receiving special education services in spring 2002. Results indicate that children receiving special education services scored about one standard deviation lower on the Reading Test in spring 2004 than those children not receiving such services (i.e., 122.9 vs. 145.2 points, respectively). This difference is statistically significant at the p <.01 level. However, children who received special education services also displayed 2.7 points more in skills gain between the springs of 2002 and 2004 than children who did not receive such services (i.e., an average gain of 23.3 vs. 20.6 points, respectively). Table 3's second and third columns display results using regression analysis. These results indicate that statistically controlling for differences in children's backgrounds substantially reduces special education's negative effect from −22.3 to −6.6 points, although it remains statistically significant at the p <.001 level. Accounting for these covariates also reduces the difference in 2002–2004 test score gains to statistical non-significance.

Results in Table 3's last three columns display the predicted effects of special education services on children's reading skills, as measured both by the children's 2004 Reading Test IRT scores and the gain between their 2002 and 2004 IRT scores. When matched on their propensity to receive special education services, children who received these services displayed neither greater reading skills in 2004 nor greater gains in their reading skills between 2002 and 2004 when compared to closely matched peers who did not receive these services. Instead, children who received special education services displayed lower reading skills. Special education's negative effect, as measured using either stratification matching (i.e., −4.0 points) or kernel matching (i.e., −3.5 points), on the children's 2004 IRT scores is statistically significant at the p <.05 to p < .01 levels. Differences in the children's gains on the 2002 to 2004 administrations of the Reading Test were again statistically non-significant.

Table 4 displays estimates of the effect of special education for children who, within each of the eight propensity strata, were predicted to have statistically similar propensities of receiving special education services. Regardless of a child's propensity to receive special education services, delivery of such services did not result in statistically significant differences in his or her reading skills.

Effects on Mathematics Skills—Table 2 displays simple mean differences on the Mathematics Test between children who were or were not placed into special education. These simple mean differences indicate that, in 2004, children who received special education services scored 16.5 points lower on the Mathematics Test than children not so placed (i.e., 101.8 vs. 118.3 points). This difference is statistically significant at the p <.01 level. The difference in the two groups' gains between 2002 and 2004 were not statistically significant. Table 3's regression analyses indicate that children placed in special education did less well in 2004 on the Mathematics Test than those children not so placed. Although adjusting for the model's covariates substantially reduces special education's negative effect from 16.4 points to 3.5 points, the estimated effect remains statistically significant at the p <.001 level.

The last three columns of Table 3 indicate that once children are matched on their propensity for special education placement, the negative effects of special education are reduced to statistical non-significance, as measured either by the 2004 IRT score or 2002 to 2004 IRT score gain. Table 4's within stratum contrasts indicate that, regardless of the child's propensity to receive special education services, delivery of such services did not lead to statistically significant differences in his or her mathematics skills.

Do Special Education Services Improve Children's Behavior?

Effects on Learning-Related Behaviors—Table 2 displays simple mean differences in the learning-related behaviors of children who received and who did not receive special education services. The mean difference of .04 indicates that teachers rated those children receiving special education services as less task-focused than children not receiving such services (i.e., 2.8 vs. 3.2 points). However, those who received special education did begin to engage in learning-related behaviors with increasing frequency in contrast to those who did not receive these services. This is indicated by the positive and statistically significant .11 point difference in the 2002–2004 gain scores in the teacher ratings (i.e., .10 points vs. −.01 points). Table 5's regression coefficients indicate that these effects are reduced to statistical non-significance after controlling for the model's covariates.

The final three columns of Table 5 indicate that once children are matched on their propensity for receiving special education services, the effect of special education on learning-related behaviors in 2004 is reduced to statistical non-significance. However, when measured as the difference in gains between children's scores on the 2002 and 2004 teacher ratings, special education has a statistically significant, positive effect on learning-related behaviors. We obtained this ATT with stratification matching, nearest neighbor matching, and kernel matching. These methods estimated that the effect of special education on increasing children's learning-related behaviors ranged between .09 and .14 of a point on the 4-point scale used by teachers to rate such behaviors.

Table 6 provides additional detail on the pattern of this positive predicted effect. Special education improves the learning-related behaviors of those children who are most likely to receive such services (i.e., strata 6, 7, and 8) when measured as the differences in teacher ratings in 2004 (i.e., *d*s of .03, .25, and 1.56, respectively, each $p<.05$). Yet it has statistically significant negative predicted effects on the behaviors of those children who are relatively less likely to receive special education services (i.e., strata 1, 2, 3, 4, and 5). The negative effect sizes range from −.07 to −.45 for these five strata. Special education has a statistically significant positive effect on the learning-related behaviors of children in strata 5–8 when the dependent variable is the differences in the gains between the teacher ratings from 2002 to 2004 (i.e., *d*s of .11, .19, .44, .47, respectively). Thus, and even after using gain scores to better control for a child's initial level of learning-related behaviors in 2002, receipt of special education services has an estimated positive effect on the learning-related behaviors of those most likely to receive such services. These positive effect sizes are approximately linear. This occurs whether they are measured using only the 2004 ratings or gain scores in the 2002–2004 ratings. However, special education yields a consistently negative (non-linear) effect on the frequency of learning-related behaviors for those children less likely to be so placed (i.e., strata 1, 2, and 3). This occurs whether the effect sizes are measured using the 2004 ratings or gain scores in the 2002–2004 ratings.

Effects on Externalizing Problem Behaviors—Table 2's mean differences indicate that children receiving special education services were rated as displaying externalizing problem behaviors more frequently than children not in special education (i.e., 1.76 versus 1.60, $p<.01$). However, no statistically significant difference is evident between the two groups of children's 2002–2004 gain scores. Table 5's regression coefficients show that special education has no statistically significant effects on the frequency of children's externalizing problem behaviors after controlling for the model's covariates.

Table 5's last three columns indicate that use of propensity score matching further reduces the estimated ATTs, and again indicate that special education services yields no statistically significant effects. Stratum-level contrasts displayed in Table 6 show that special education consistently reduces the frequency with which children who are relatively likely to be in special education (i.e., stratum 6, 7, & 8) engage in externalizing problem behaviors. However, this linear effect is limited to analyses using the children's 2004 scores on the teacher rating. A less consistent pattern is found when analyzing the children's 2002 to 2004 gain scores. Here, only those children in stratum 6 display less frequent externalizing problem behaviors, although the size of the statistically significant effect size is very small (i.e., *d* of −.04). Children in strata 1, 3, and 7 now display more frequent externalizing behaviors as a function of receiving special education services. However, the effect sizes for these statistically significant positive differences in gain scores are again small, ranging from .09 to .11.

Effects on Internalizing Problem Behaviors—The mean differences in Table 2 indicate that children who were in special education in spring 2002 displayed internalizing problem behaviors more frequently in spring 2004 than children who had not been so placed (i.e., 1.88 vs. 1.59, $p<.01$). However, this effect is not evident when using 2002 to 2004 gain scores. The regression analyses results displayed in Table 5 show the same pattern. The 2004 coefficients remain statistically significant even after accounting for the child's background characteristics (i.e., a reduction to .112 points, $p<.001$). However, the coefficient for the difference in the two groups of children's 2002–2004 gains is statistically non-significant whether these covariates are accounted for or not.

Results displayed in Table 5 indicate that use of propensity score matching further reduces the size of the estimated 2004 ATTs. The gains from 2002 to 2004 are not statistically significant. Thus, and as an overall point estimate, receipt of special education services does not decrease the frequency of children's internalizing problem behaviors. The within-stratum contrasts of Table 6 show greater detail underlying this point estimate. These contrasts indicate that special education increases the frequency of most children's (i.e., those in strata 1, 2, 3, 4, 5, 6, but not 7) internalizing problem behaviors. Yet this predicted effect is only evident using the children's 2004 teacher rating scores. These statistically significant effects are mostly small (i.e., *d*s of .36, .06, .26, .26, .12, .20, −.35, respectively). The 2002–2004 gain scores yield inconsistent predicted effects. Here, special education increases the frequency of the child's internalizing problem behaviors of children in strata 1 *(d* of .53)*,* but decreases the frequency of such behaviors of children in strata 3 and 5 (*d*s of −.11 and −.14, respectively).

Discussion

We sought to quantify the effects of naturally delivered special education services on U. S. schoolchildren's learning and behavior. We did so using a large and longitudinal sample of such children and methods that should greatly reduce selection bias. Because we could not randomly assign children to receive or not receive special education services, we used propensity score matching techniques to contrast the learning and behavior of children who did and who did not receive special education services, but who were matched on a wide range of observed background characteristics. We also conducted "benchmark" analyses using *t* tests of mean differences and OLS regression.

Evidence of Special Education's Effects

It is widely considered critical that children acquire basic skills proficiency if they are to experience long-term educational and societal opportunities (e.g., Reynolds, Elksnin, & Brown, 1996; Snow, Burns, & Griffin, 1998; U.S. Department of Education, 2007). For example, a lack of proficiency in mathematics lowers an adult's employability and wages, over and above poor reading ability, low IQ, and many other factors (Rivera-Batiz, 1992). Even those adults with good literacy skills are more likely to be unemployed and less likely to be promoted when employed if they have poor mathematics skills (Parsons & Bynner, 1997). Yet a large majority of high school youth with disabilities displays below-basic levels of proficiency in both mathematics and reading (NAEP, 2005).

Our analyses indicate that special education services being provided to U.S. schoolchildren during their elementary years may not be of sufficient strength to prevent a subsequent lack of basic skills proficiency. Specifically, we found that special education services had negative or statistically non-significant effects on young children's reading and mathematics skills. We found that children receiving special education services in the spring of 2002 displayed significantly lower reading skills in the spring of 2004 than closely matched peers not receiving such services. These two groups of children displayed statistically equivalent reading skills gain between 2002 and 2004. The lack of statistically significant effects on mathematics learning was evident whether measured as either the difference between the two groups of children's 2004 IRT scores or as the difference in gain between their 2002 and 2004 IRT scores. Regardless of a child's propensity to receive special education services, receipt of such services resulted in no statistically significant change in his or her reading or mathematics skills.

Special education also had negative or statistically non-significant effects on most children's externalizing or internalizing problem behaviors. That is, overall, special education services failed to lessen the frequency with which children engaged in these behaviors. This is problematic because children who display such behaviors frequently are at greater risk for a range of negative long-term outcomes (e.g., Schaeffer et al., 2003; Sprague & Walker, 2000). These behaviors can also quickly become resistant to intervention, necessitating that they be prevented or remediated by the elementary grades (e.g., Walker, Colvin, & Ramsey, 1995). However, we did find that these estimates varied by a child's propensity to receive special education services. The receipt of special education services reduced the frequency with which children who were relatively likely to receive special education services engaged in externalizing problem behaviors. This pattern was generally limited to those analyses using the children's 2004 teacher rating scores. Within-stratum contrasts indicated that receipt of special education increased the frequency of some children's internalizing problem behaviors. This predicted effect was again generally limited to the children's 2004 teacher rating scores.

We also found that receipt of special education services had positive predicted effects on children's learning-related behaviors. This is an important finding, which to our knowledge has not been previously reported. Learning-related behavior is increasingly shown to be a key contributor to young children's skills proficiency in reading (e.g., Tach & Farkas, 2006) and mathematics (e.g., Diperna et al., 2007). Thus, it may be that, over time, special education services positively but indirectly impact both types of skills. Yet the magnitude of this indirect effect may be small. Our study's largest positive ATT was only .14 of one point's difference in gain in children's learning-related behaviors.

Special education is sometimes hypothesized to have stigmatizing effects on children, such that children so placed may be more likely to display task-avoidant, acting out, or social withdrawal behaviors (La Greca & Stone, 1990; Valas, 2001). Our estimates of special education's ATT on children's externalizing or internalizing problem behaviors fail to provide consistent evidence for this hypothesis. However, our within-stratum contrasts do repeatedly indicate that children who were predicted to be unlikely to be placed into special education, but who nevertheless were so placed (i.e., those in stratum 1), more frequently displayed these two types of problem behaviors, as well as less frequent learning-related behaviors. If special education placement has stigmatizing effects on children, then these effects are most likely experienced by children whose background characteristics are most like those of children who are not so placed.

We derived these estimates of special education's effectiveness using rigorous methodology. We used multiple measures of children's learning and behavior. Each of these measures displays strong psychometric properties. Our sample was selected from a large-scale, longitudinal, and nationally representative sample of U.S. elementary schoolchildren. We also estimated special education's impact using multiple analytical methods, including three different types of propensity score matching techniques. Collectively, these methods yielded the same general pattern of findings, which helps ensure that flawed methodology is an unlikely explanation for the study's findings.

Limitations

Nevertheless, our study has several limitations. Our propensity score model includes factors identified in prior research as predictive of the receipt of special education services. However, our model may not have included additional factors that predict a child's receipt of such services. Therefore our results may still be affected by hidden (i.e., omitted variable) bias. Children in our "control" group were not formally referred, evaluated, and identified as disabled. Instead, their observed background characteristics were very similar to those of children not so identified. Propensity score matching can only approximate randomization in the identification process. However, we did match children on factors (e.g., a child's relative proficiency in reading and mathematics) that have been identified in prior research as being most predictive of whether or not a low performing child is identified as disabled by a multidisciplinary team (Merrell & Shinn, 1990). We were unable to account for the multi-level structure of the data (i.e., students nested within schools). Doing so was not a viable analytical strategy because only 7% of the sample's schools were attended by more than one student receiving special education services. Our study was designed to provide a general or overall estimate of special education's effects. Our intent-to-treat analyses provide estimates of special education's "use-effectiveness," rather than its "method-effectiveness." We can currently offer no detail on the effectiveness of specific types of special education services. For example, we are unable to say whether special education services were more or less effective when delivered in traditional rather than inclusionary settings. Further investigations of the effects of particular types of special education services are clearly warranted. We are also unable to say whether special education services are more effective when used with children with particular types of disabilities (e.g., mental retardation vs. learning disabilities). Our point estimates of special education's effects are limited to children's elementary school years. Analyses across a longer time period may have yielded more positive effects. Our study also did not measure additional types of learning (e.g., general knowledge) or behavior (e.g., interpersonal skills) that special education services may impact. We used teacher ratings to estimate special education's effects on children's behaviors. Yet some teachers may provide biased reports of children's behavior (Taylor, Gunter, & Slate, 2001). Our estimates for behavior therefore may be confounded by teacher bias. These behavioral estimates are also limited to the effects of special education services in only one context (i.e., school). Gain scores can have problematic psychometric properties (e.g., low reliability ([Cronbach & Furby, 1970]). They may also yield biased estimates of a treatment effect (e.g., due to a lack of stationarity in the true scores ([Maris, 1998]). However, use of IRT-scaled scores when calculating gain scores (as was the case here for the ECLS-K's Reading and Mathematics Tests) helps avoid these psychometric limitations (Wang & Chyi-In, 2004).

We also acknowledge that the study's results could be characterized as "glass half full" evidence of special education's effects. This is because the learning and behavior of children receiving special education services were often statistically equivalent to that of matched children not receiving such services. Those receiving special education services displayed about the same level of reading skill in 2004 (i.e., an average IRT Reading Test score of 122.91) as that displayed in 2002 by closely matched children not receiving such services (i.e., an average IRT Reading Test score of 124.63). However, and again, our contrasts involved two groups of children who were closely matched, as they displayed the same or nearly the same propensities to be placed into special education. We therefore characterize the estimates as "glass half empty" because they indicate that special education has mostly negative or statistically non-significant effects on children's learning and behavior, despite these services' very high costs. Put another way, our study finds little evidence that special education services, as currently being implemented in U. S. schools, are positively impacting the learning or behavior of most children with disabilities, despite the vast resources being invested in the provision of such services.

Implications of the Study's Results

A range of practices has been empirically shown to positively impact the learning and behavior of children with disabilities (e.g., Forness, Freeman, & Paparella, 2006; Fuchs, Fuchs, Mathes, & Simmons, 1997; Howlin, Gordon, Pasco, Wade, & Charman, 2007; Morgan & Sideridis, 2006; Schwartz, Carta, & Grant, 1996; Stecker, Fuchs, & Fuchs, 2005; Swanson & Hoskyn, 1998). Yet classroom observation studies repeatedly indicate that children do not always receive special education services that can reasonably be expected to mitigate the effects of their disabilities (e.g., Lane et al., 2005; Levy & Vaughn, 2002). For example, Harry and Klingner's (2006) observations indicate that special education programs were often "marked by routine and generic, rather than individualized, instruction; teacher shortages; widely variable teacher quality; unduly restrictive environments in some programs…; and unduly large class sizes" (p. 172). Others have noted the wide-spread shortages of qualified teachers (McLeskey, Tyler, & Flippin, 2004), ever increasing numbers of children being placed into special education (Russ, Chiang, Rylance, & Bongers, 2001; U.S. Department of Education, 2005), and infrequent use of research-based practices by school staff (Boardman, Arguelles, Vaughn, Hughes, & Klingner, 2005). Policy-makers, researchers, and practitioners have repeatedly worked to increase the effectiveness of special education services (e.g., Denston, Vaughn, & Fletcher, 2003; Deshler, 2005; Gersten et al., 1997). To date, however, the empirical studies indicating the need for such efforts have been methodologically limited and have not directly estimated how extensively special education services in the U.S. may be failing to improve the learning or behavior of children with disabilities. Our study's findings add much urgency to efforts to increase the effectiveness of special education services being delivered in U.S schools.

Acknowledgments

This research was supported by an American Educational Research Association Research Grant (George Farkas & Paul Morgan, Co-Principal Investigators) as well as a grant from the National Institute of Child Health and Human Development (R01-HD050144; Gary Sandefur, Principal Investigator). Opinions reflect those of the authors and not necessarily those of the granting agencies. We thank Kenneth Frank, Michael Massoglia, John Lloyd, and Kevin Sutherland for their helpful comments.

References

Abbott M, Walton C, Tapia Y, Greenwood C. Research to practice: A "blueprint" for closing the gap in local schools. Exceptional Children. 1999; 65:339–352.

Allison PD. Change scores as dependent variables in regression analysis. Sociological Methodology. 1990; 20:93–114.

Andrews H, Goldberg D, Wellen N, Pittman B, Struening E. Prediction of special education placement from birth certificate data. American Journal of Preventative Medicine. 1995; 11:55–61.

Bateman, BD.; Linden, MA. Better IEPs: How to develop legally correct and educationally useful programs. 4. Verona, WI: IEP Resources, Attainment Co; 2006.

Becker SO, Ichino A. Estimation of average treatment effects based on propensity scores. Stata Journal. 2002; 2:358–377.

Blackorby J, Wagner M. Longitudinal postschool outcomes of youth with disabilities: Findings from the National Longitudinal Transition Study. Exceptional Children. 1996; 62:399–413.

Boardman AG, Arguelles ME, Vaughn S, Hughes MT, Klingner J. Special education teachers' views of research-based practices. The Journal of Special Education. 2005; 39:168–180.

Carnine D. Bridging the research-to-practice gap. Exceptional Children. 1997; 63:513–521.

Cronbach LJ, Furby L. How we should measure change or should we? Psychological Bulletin. 1970; 74:68–80.

Czajka JL, Hirabayashi SM, Little RJA, Rubin DB. Projecting from advance data using propensity modeling: An application to income and tax statistics. Journal of Business and Economic Statistics. 1992; 10:117–131.

Dehejia RH, Wahba S. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. Journal of the American Statistical Association. 1999; 94:1053–1062.

Delgado CEF, Vagi SJ, Scott KG. Early risk factors for speech and language impairments. Exceptionality. 2005; 13:173–191.

Denton CA, Vaughn S, Fletcher JM. Bringing research-based practice in reading intervention to scale. Learning Disabilities Research & Practice. 2003; 18:201–211.

Deshler, DD. Intervention research and bridging the gap between research and practice. Electronically. 2005. retrieved on August, 8, 2007 at http://www.ldonline.org/article/5596

DiPerna JC, Lei PW, Reid EE. Kindergarten predictors of mathematical growth in the primary grades: An investigation using the Early Childhood Longitudinal Study—Kindergarten Cohort. Journal of Educational Psychology. 2007; 99:396–379.

Donovan, MS.; Cross, CT. Minority Students in Special and Gifted Education. Washington, DC: National Academy Press; 2002.

Feng H, Cartledge G. Social skill assessment of inner city Asian, African, and European American students. School Psychology Review. 1996; 25:228–239.

Fergusson DM, Woodward LJ. Mental health, educational, and social role outcomes of adolescents with depression. Archives of General Psychiatry. 2002; 59:225–231. [PubMed: 11879160]

Forness SR, Freeman SFN, Paparella T. Recent randomized clinical trials comparing behavioral interventions and psychopharmacologic treatments for students with EBD. Behavioral Disorders. 2006; 31:284–296.

Frisco ML, Muller C, Frank K. Family structure change and adolescents' school performance: A propensity score approach. Journal of Marriage and Family. 2007; 69:721–741. [PubMed: 20300482]

Fuchs D, Fuchs LS, Mathes PG, Simmons DC. Peer-Assisted Learning Strategies: Making classrooms more responsive to diversity. American Educational Research Journal. 1997; 34:174–206.

Furlong, MJ.; Karno, M. Review of the Social Skills Rating System. In: Conoley, JC.; Impara, JC., editors. Twelfth mental measurements yearbook. Lincoln, NE: Buros Institute of Mental Measurement; 1995. p. 697-969.

Gersten R, Baker S, Lloyd JW. Designing high-quality research in special education: Group experimental design. Journal of Special Education. 2000; 34:2–18.

Gersten R, Vaughn S, Deshler D, Schiller E. What we know about using research findings: Implications for improving special education practice. Journal of Learning Disabilities. 1997; 30:466–476. [PubMed: 9293227]

Goodman, LA. Analyzing Qualitative/Cetegorical Data. Cambridge, MA: Abt Books; 1978.

Green KM, Ensminger ME. Adult social behavioral effects of heavy adolescent marijuana use among African Americans. Developmental Psychology. 2006; 42:1168–1178. [PubMed: 17087550]

Gresham, PM.; Elliot, SN. Social Skills Rating System. Circle Pines, MN: American Guidance Service; 1990.

Harry, B.; Klingner, JK. Why are so many minority students in special education: Understanding race and disability in schools. New York: Teachers College Press; 2006.

Harknett K. Does receiving an earnings supplement affect union formation? Estimating effects for program participants using propensity score matching. Evaluation Review. 2006; 30:741–778. [PubMed: 17093107]

Heckman JJ. Identification of causal effects using instrumental variables: Comment. Journal of the American Statistical Association. 1996; 91:459–462.

Heckman JJ, Ichimura H, Todd PE. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. Review of Economic Studies. 1997; 64:605–654.

Hibel, J.; Farkas, G.; Morgan, PL. Who is placed into special education?. 2008. Manuscript under review

Hollomon HA, Dobbins DR, Scott KG. The effects of biological and social risk factors on special education placement: Birthweight and maternal education as an example. Research in Developmental Disabilities. 1998; 19:281–294. [PubMed: 9653804]

Hong G, Raudenbush SW. Effects of Kindergarten retention policy on children's cognitive growth in reading and mathematics. Educational Evaluation and Policy Analysis. 2005; 27:205–224.

Hosp JL, Reschly DJ. Predictors of restrictiveness of placement for African American and Caucasian students with learning disabilities. Exceptional Children. 2002; 68:225–238.

Hosp JL, Reschly DJ. Disproportionate representation of minority students in special education: Academic, demographic, and economic predictors. Exceptional Children. 2004; 70:185–199.

Howlin P, Gordon RK, Pasco G, Wade A, Charman T. The effectiveness of Picture Exchange Communication System (PECS) training for teachers of children with autism: A pragmatic, group randomised controlled trial. Journal of Child Psychology and Psychiatry. 2007; 48:473–481. [PubMed: 17501728]

Individuals with Disabilities Improvement Education Act, Public Law 108–446, Statute number 2651 (2004).

Jones AS, D'Agostino RB, Gondolf EW, Heckert A. Assessing the effect of batterer program completion on re-assault using propensity scores. Journal of Interpersonal Violence. 2004; 19:1002–1020. [PubMed: 15296614]

Kansas State Legislative Research Department. Estimated special education excess costs—FY 2008. Electronically. 2004. retrieved on August 8, 2007 at http://www.ksde.org/LinkClick.aspx?fileticket=lv0CzOcvet0%3D&tabid=119&mid=4087

Katusic SK, Colligan RC, Barbaresi WJ, Schaid DJ, Jacobsen SJ. Incidence of reading disability in a population-based birth cohort, 1976–1982, Rochester, Minn. Mayo Clinic Proceedings. 2001; 76:1081–1092. [PubMed: 11702896]

Kavale, KA. Learning disability and cultural-economic disadvantage: The case for a. 1988.

La Greca AM, Stone WL. LD status and achievement: Confounding variables in the study of children's social status, self-esteem, and behavioral functioning. Journal of Learning Disabilities. 1990; 23:483–490. [PubMed: 2246600]

Lachin JM. Statistical considerations in the intent-to-treat principle. Controlled Clinical Trials. 2000; 21:241–243. [PubMed: 10822121]

Lane KL, Wehby JH, Little MA, Cooley C. Students educated in self-contained classrooms and self-contained schools: Part II. How do they progress over time. Behavioral Disorders. 2005; 30:363–374.

Lechner M. Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies. Review of Economics and Statistics. 2002; 84:205–220.

Levy S, Vaughn S. An observational study of reading instruction of teachers of students with emotional/behavioral disorders. Behavioral Disorders. 2002; 27(3):215–235.

Luellen JK, Shadish WR, Clark MH. Propensity scores: An introduction and experimental test. Evaluation Review. 2005; 29:530–58. [PubMed: 16244051]

Magiera K, Zigmond N. Co-teaching in middle school classrooms under routine conditions: Does the instructional experience differ for students with disabilities in co-taught and solo-taught classes? Learning Disabilities Research & Practice. 2005; 20:79–85.

Mann EA, McCartney K, Park JM. Preschool predictors of the need for early remedial and special education services. Elementary School Journal. 2007; 107:273–285.

Mannerkoski MK, Aberg LE, Auti TH, Hoikkala M, Sarna S, Heiskala HJ. Newborns at risk for special-education placement: A population-based study. European Journal of Paediatric Neurology. 2007; 11:223–231. [PubMed: 17346999]

Maris E. Covariance adjustment versus gain scores—Revisited. Psychological Methods. 1998; 3:309–327.

McClelland MM, Acock AC, Morrison FJ. The impact of kindergarten learning-related skills on academic trajectories at the end of elementary school. Early Childhood Research Quarterly. 2006; 21:471–490.

McLeskey J, Tyler NC, Flippin SS. The supply of and demand for special education teachers: A review of research regarding the chronic shortage of special education teachers. Journal of Special Education. 2004; 38:5–21.

Meisels, SJ.; Atkins-Burnett, S.; Nicholson, J. Assessment of social competence, adaptive behaviors, and approaches to learning with young children. Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement; 1996.

Merrell KW, Shinn MR. Critical variables in the learning disabilities identification process. School Psychology Review. 1990; 19:74–90.

Morgan PL, Sideridis GD. Contrasting the effectiveness of fluency interventions for students with or at risk for learning disabilities: A multilevel random coefficient modeling meta-analysis. Learning Disabilities: Research and Practice. 2006; 21:191–210.

National Center for Education Statistics. Early Childhood Longitudinal Study-Kindergarten Class of 1998–1999 (ECLS-K): Psychometric report for kindergarten through first grade (Working Paper Number: 2002–05). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement; 2002.

National Center for Education Statistics. User's manual for the ECLS-K third grade public-use data file and electronic code book (NCES 2004–001). Washington, DC: U.S. Department of Education, Institute for Education Sciences; 2004.

National Center for Education Statistics. Early Childhood Longitudinal Study, Kindergarten Class of 1998–1999 (ECLS-K): Psychometric report for the third grade (NCES 2005–062). Washington, DC: U.S. Department of Education, Institute of Education Sciences; 2005.

Pagan, A.; Ullah, A. Nonparametric econometrics. Cambridge University Press; 1999.

Parsons S, Bynner J. Numeracy and employment. Education and Training. 1997; 39:43–51.

Petkova E, Teresi J. Some statistical issues in the analyses of data from longitudinal studies of elderly chronic care populations. Psychosomatic Medicine. 2002; 64:531–547. [PubMed: 12021427]

Phelps A, Hanley-Maxwell C. School to work transitions for youth with disabilities: A review of outcomes and practices. Review of Educational Research. 1997; 67:197–226.

President's Commission on Excellence in Special Education. A New Era: Revitalizing Special Education for Children and their Families. 2002. Retrieved December 4, 2007 at http://www.ed.gov/inits/commissionsboards/whspecialeducation/reports/index.html

Reynolds AM, Elksnin N, Brown FR. Specific reading disabilities: Early identification and long-term outcome. Mental Retardation and Developmental Disabilities Research Reviews. 1996; 2:21–27.

Rivera-Batiz FL. Quantitative literacy and the likelihood of employment among young adults in the United States. The Journal of Human Resources. 1992; 27:313–328.

Rosenbaum PR. Dropping out of high school in the United States: An observational study. Journal of Educational Statistics. 1986; 11:207–224.

Rosenbaum PR, Rubin D. The central role of the propensity score in observational studies of causal effects. Biometrika. 1983; 70:41–55.

Rubin DB. Estimating causal effects from large data sets using propensity scores. Annals of Internal Medicine. 1997; 127:757–763. [PubMed: 9382394]

Russ S, Chiang B, Rylance BJ, Bongers J. Caseload in special education: An integration of research findings. Exceptional Children. 2001; 67:161–172.

Schaefer C, Petras H, Ialongo N, Poduska J, Kellam S. Modeling growth in boys' aggressive behavior across elementary school: Links to later criminal involvement, conduct disorder, and anti-social personality disorder. Developmental Psychology. 2003; 39:1020–1035. [PubMed: 14584982]

Schneider B. Educational stratification and the life course. Sociological Focus. 2001; 34:463–466.

Schneider, B., et al. Estimating causal effects using experimental and observational designs. Washington, D.C: American Educational Research Association; 2007.

Schwartz I, Carta J, Grant S. Examining the use of recommended language intervention practices in early childhood special education classrooms. Topics in Early Childhood Special Education. 1996; 16:251–72.

Shadish, WR.; Cook, TD.; Campbell, DT. Experimental and quasi-experimental design for generalized causal inference. Boston: Houghton-Mifflin; 2002.

Shores RE, Jack SL, Gunter PL, Ellis DN, DeBriere TJ, Wehby JH. Classroom interactions of children with behavior disorders. Journal of Emotional & Behavioral Disorders. 1993; 1:27–39.

Snell ME. Applying research to practice: The more pervasive problem? Research & Practice for Persons with Severe Disabilities. 2003; 28:143–147.

Snow, CE.; Burns, MS.; Griffin, P. Preventing reading difficulties in young children. Washington, DC: National Academic; 1998.

Sprangue J, Walker H. Early identification and intervention for youth with antisocial and violent behavior. Exceptional Children. 2000; 66:367–379.

Stanovich K. Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. Reading Research Quarterly. 1986; 21:360–407.

Stecker PM, Fuchs LS, Fuchs D. Using curriculum-based measurement to improve student achievement: Review of research. Psychology in the Schools. 2005; 42:795–819.

Stone RA, Obrosky DS, Singer DE, Kapoor WN, Fine MJ. the Pneumonia Patient Outcomes Research Team (PORT) Investigators. Propensity score adjustment for pretreatment differences between hospitalized and ambulatory patients with community-acquired pneumonia. Medical Care. 1995; 33:56–66.

Swanson HL, Hoskyn M. Experimental intervention research on students with learning disabilities: A meta-analysis of the treatment outcomes. Review of Educational Research. 1998; 68:277–321.

Tach L, Farkas G. Learning-related behaviors, cognitive skills, and ability grouping when schooling begins. Social Science Research. 2006; 35:1048–1079.

Taylor PB, Gunter PL, Slate JR. Teachers' perceptions of inappropriate student behavior as a function of teachers' and students' gender and ethnic background. Behavioral Disorders. 2001; 26:146–15.

U. S. Department of Education. Twenty-fourth annual report to Congress on the implementation of the Individuals with Disabilities Education Act. Washington DC: U.S. Government Printing Office; 2002.

U. S. Department of Education. 27th Annual Report to Congress on the Implementation of the Individuals with Disabilities Education Act. 2005; 1 Retrieved December 3, 2007 at http://www.ed.gov/about/reports/annual/osep/2005/parts-b-c/27th-vol-1.pdf.

U.S. Department of Education, Office of Special Education Programs. IDEA regulations: Individualized Education Program (IEP). 2006. Retrieved August 10, 2007 at http://idea.ed.gov/explore/view/p/%2Croot%2Cdynamic%2CTopicalBrief%2C10%2C

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. Results. 2007. retrieved July 18, 2007 at http://nces.ed.gov/nationsreportcard/

U.S. Department of Education. The Math Now Programs. 2007. Retrieved August 8, 2007 at http://www.ed.gov/news/pressreleases/2006/02/02062006.html

U. S. Department of Education, National Center for Education Statistics. Demographic and school characteristics of students receiving special education in the elementary grades. 2007. (NCES 2007–005). Retrieved December 3, 2007 at http://nces.ed.gov/pubs2007/2007005.pdf

Valas H. Learned helplessness and psychological adjustment II: Effects of learning disabilities and low achievement. Scandinavian Journal of Educational Research. 2001; 45:101–114.

Walker, HM.; Colvin, G.; Ramsey, E. Antisocial behavior in school: Strategies and best practices. Pacific Grove, CA: Brooks/Cole Publishing Company; 1995.

Wang WC, Chyi-In W. Gain score in Item Response Theory as an effect size measure. Educational and Psychological Measurement. 2004; 64:758–780.

Williams RH, Zimmerman DW. Are simple gain scores obsolete? Applied Psychological Measurement. 1996; 20:59–69.

Winship C, Mare RD. Models for sample selection bias. Annual Review of Sociology. 1992; 18:327–350.

Table 1

Balance of the Propensity Scores for Spring, 2002 Special Education Placement

	Special Education Students			Non-Special Education Students		
Stratum	***N***	***M***	***SD***	***N***	***M***	***SD***
1	27	0.007	0.005	3,558	0.006	0.004
2	24	0.021	0.004	1,283	0.023	0.005
3	42	0.045	0.009	1,118	0.044	0.009
4	67	0.088	0.160	785	0.089	0.018
5	110	0.181	0.036	470	0.174	0.034
6	141	0.363	0.070	220	0.348	0.068
7	77	0.621	0.079	52	0.596	0.071
8	38	0.862	0.063	8	0.838	0.061

Table 2

Dependent Variable Means and Standard Deviations for Analytical Sample, Special Education Students, and Non-Special Education Students

Dependent Variables	**Full Sample (*N*=6,318)**	**Special Education Students (*n*=363)**	**Non-Special Education Students (*n*=5,955)**
Reading IRT Test	143.90 (20.94)	122.91 ** (24.54)	145.18 ** (19.99)
Reading IRT Test Score Gain	20.71 (12.40)	23.32 ** (14.17)	20.55 ** (12.26)
Math IRT Test	117.30 (19.55)	101.83 ** (23.44)	118.25 ** (18.88)
Math IRT Test Score Gain	21.23 (10.36)	20.90 (11.41)	21.25 (10.32)
Approaches to Learning	3.12 (0.66)	2.76 ** (0.63)	3.15 ** (0.65)
Approaches to Learning Gain	0.00 (0.62)	0.10 ** (0.62)	−0.01 ** (0.62)
Externalizing Behaviors	1.61 (0.55)	1.76 ** (0.62)	1.60 ** (0.55)
Externalizing Behaviors Gain	−0.04 (0.55)	−0.06 (0.58)	−0.04 (0.55)
Internalizing Behaviors	1.61 (0.53)	1.88 ** (0.60)	1.59 ** (0.52)
Internalizing Behaviors Gain	0.02 (0.62)	0.02 (0.68)	0.02 (0.61)

Note:

** Means of special education and non-special education students are significantly different at the $p < .01$ level; *SD* in parentheses.

Table 3

Estimated Effects of Special Education Services on Children’s Reading and Mathematics Skills

	OLS Regression (no covariates)	**OLS Regression (covariates)**	**Stratification Matching**	**Nearest Neighbor Matching**	**Kernel Matching**
2004 Reading IRT Test scores	−22.276***	−6.622***	−3.995**	−2.872	−3.514*
2002–2004 Reading IRT Test score gains	2.772***	.789	1.080	.955	.802
2004 grade IRT Math Test scores	−16.415***	−3.350***	−1.732	−2.161	−1.704
2002–2004 IRT Math Test score gains	−.351	−.693	−.574	−.617	−.389

* *p*<.05;

** *p*<.01;

*** *p*<.001

Table 4

Within-Stratum Differences in Reading and Mathematics Skills between Special Education and Non-Special Education Students

	Special Education Students			Non-Special Education Students			
Stratum	***N***	***M***	***SD***	***N***	***M***	***SD***	***Effect Size (d)***
	2004 Reading Test Score						
1	18	160.68	11.06	2873	153.91	15.72	0.43
2	19	145.99	21.87	996	142.62	17.50	0.19
3	30	134.19	16.74	889	140.12	18.43	−0.32
4	51	123.22	20.79	624	133.37	20.34	−0.50
5	83	123.75	21.87	371	129.15	20.91	−0.26
6	94	118.05	22.92	169	123.70	22.07	−0.26
7	47	112.71	17.54	29	117.94	24.38	−0.21
8	21	94.02	23.49	4	80.49	4.97	2.72
	2002–2004 Reading Test Score Gains						
1	18	21.13	11.82	2873	19.25	11.44	0.16
2	19	21.99	15.84	996	20.69	12.35	0.11
3	30	23.25	15.78	889	21.40	12.54	0.15
4	51	20.49	13.98	624	22.37	13.13	−0.14
5	83	24.40	11.67	371	23.63	14.13	0.05
6	94	24.22	14.68	169	24.01	13.25	0.02
7	47	25.83	13.91	29	20.26	12.19	0.46
8	21	19.47	19.17	4	18.12	4.11	0.33
	2004 Math Test Score						
1	18	131.91	11.08	2873	125.11	14.83	0.46
2	19	119.84	17.10	996	116.58	18.09	0.18
3	30	112.47	18.18	889	114.07	18.07	−0.09
4	51	104.81	21.77	624	109.29	20.40	−0.22
5	83	102.83	20.96	371	105.08	20.97	−0.11
6	94	98.55	21.70	169	100.17	23.50	−0.07
7	47	87.40	21.89	29	96.94	22.04	−0.43
8	21	80.35	21.91	4	68.39	14.16	0.84
	2002 – 2004 Math Test Score Gains						

	Special Education Students			Non-Special Education Students			
Stratum	***N***	***M***	***SD***	***N***	***M***	***SD***	***Effect Size (d)***
1	18	20.37	11.32	2873	20.61	9.73	−0.02
2	19	20.82	10.29	996	21.51	10.24	−0.07
3	30	23.44	8.08	889	22.25	10.90	0.11
4	51	22.15	11.23	624	21.89	10.53	0.02
5	83	19.56	11.55	371	21.78	11.62	−0.19
6	94	22.33	11.33	169	21.98	12.42	0.03
7	47	18.99	12.97	29	22.10	12.36	−0.25
8	21	17.98	12.77	4	15.24	9.18	0.30

Note: None of these contrasts were statistically significant at $p<.05$

Table 5

Estimated Effects of Special Education Services on Children’s Behaviors

	OLS Regression (no covariates)	OLS Regression (covariates)	Stratification Matching	Nearest Neighbor Matching	Kernel Matching
2004 Approaches to Learning	−.389***	−.046	−.009	.027	−.008
2002–2004 Approaches to Learning gains	.109**	.052	.085*	.140*	.127***
2004 Externalizing Problem Behaviors	.164***	.013	−.031	−.090	−.039
2002–2004 Externalizing Problem Behaviors gains	−.012	.030	.005	.029	.028
2004 Internalizing Problem Behaviors	.287***	.112***	.057	.045	.093*
2002–2004 Internalizing Problem Behaviors gains	.003	.014	−.008	.029	.043

* *p*<.05;

** *p*<.01;

*** *p*<.001

Table 6

Within-Stratum Differences in Behaviors between Special Education and Non-Special Education Students

	Special Education Students			Non-Special Education Students			
Stratum	*N*	*M*	*SD*	*N*	*M*	*SD*	*Effect Size (d)*
	2004 Approaches to Learning						
1	18	3.06	0.64	2873	3.33	0.60	−0.45*
2	19	3.05	0.67	996	3.11	0.64	−0.09*
3	30	2.86	0.73	889	3.01	0.63	−0.24*
4	51	2.80	0.56	624	2.92	0.65	−0.19*
5	83	2.74	0.63	371	2.78	0.62	−0.07*
6	94	2.68	0.62	169	2.66	0.63	0.03*
7	47	2.73	0.61	29	2.58	0.56	0.25*
8	21	2.46	0.63	4	2.07	0.25	1.56*
	2002–2004 Approaches to Learning Gains						
1	18	−0.33	0.45	2873	−0.03	0.61	−0.49*
2	19	−0.02	0.47	996	0.00	0.62	−0.05*
3	30	−0.09	0.52	889	−0.02	0.63	−0.11*
4	51	0.09	0.74	624	0.07	0.64	0.03
5	83	0.13	0.59	371	0.06	0.67	0.11*
6	94	0.16	0.65	169	0.04	0.63	0.19*
7	47	0.27	0.59	29	−0.08	0.79	0.44*
8	21	0.20	0.52	4	0.00	0.43	0.47*
	2004 Externalizing Problem Behaviors						
1	18	1.56	0.42	2873	1.52	0.51	0.08*
2	19	1.62	0.58	996	1.62	0.57	0.00
3	30	1.78	0.68	889	1.64	0.55	0.26*
4	51	1.66	0.51	624	1.70	0.57	−0.07*
5	83	1.77	0.68	371	1.72	0.62	0.08*

Stratum	Special Education Students N	M	SD	Non-Special Education Students N	M	SD	Effect Size (d)
6	94	1.77	0.61	169	1.80	0.62	−0.05 *
7	47	1.87	0.65	29	1.95	0.62	−0.13 *
8	21	1.98	0.63	4	2.53	0.59	−0.93 *
	2002–2004 Externalizing Problem Behaviors Gains						
1	18	0.02	0.52	2873	−0.03	0.53	0.09 *
2	19	−0.03	0.52	996	−0.01	0.53	−0.04
3	30	0.00	0.54	889	−0.06	0.53	0.11 *
4	51	−0.11	0.47	624	−0.11	0.57	0.00
5	83	−0.09	0.62	371	−0.10	0.65	0.02
6	94	−0.09	0.65	169	−0.05	0.68	−0.04 *
7	47	0.00	0.58	29	−0.07	0.85	0.09 *
8	21	0.06	0.53	4	0.15	0.61	−0.15
	2004 Internalizing Problem Behaviors						
1	18	1.68	0.67	2873	1.51	0.47	0.36 *
2	19	1.64	0.47	996	1.62	0.53	0.06 *
3	30	1.78	0.47	889	1.63	0.53	0.26 *
4	51	1.80	0.60	624	1.66	0.54	0.26 *
5	83	1.83	0.60	371	1.75	0.59	0.12 *
6	94	2.01	0.63	169	1.88	0.65	0.20 *
7	47	1.91	0.51	29	2.18	0.78	−0.35 *
8	21	2.15	0.69	4	2.15	0.63	0.02
	2002–2004 Internalizing Problem Behaviors Gain						
1	18	0.32	0.57	2873	0.03	0.55	0.53 *
2	19	0.04	0.54	996	0.03	0.63	0.03
3	30	−0.05	0.46	889	0.02	0.62	−0.11 *
4	51	−0.02	0.63	624	−0.02	0.69	0.00
5	83	−0.14	0.75	371	−0.04	0.71	−0.14 *

	Special Education Students			Non-Special Education Students			
Stratum	***N***	***M***	***SD***	***N***	***M***	***SD***	***Effect Size (d)***
6	94	0.09	0.70	169	0.06	0.76	0.04
7	47	0.08	0.59	29	0.09	1.14	−0.01
8	21	0.11	0.91	4	0.15	0.89	−0.05

[*] p<.05