Journal of Early Intervention
2018, Vol. 40(4) 347–362

DOI: 10.1177/1053815118793430
journals.sagepub.com/home/jei
SAGE

Using Propensity Score Weighting to Reduce Selection Bias in Large-Scale Data Sets

Crystal D. Bishop[1], Walter L. Leite[1], and Patricia A. Snyder[1]

Abstract

Data sets from large-scale longitudinal surveys involving young children and families have become available for secondary analysis by researchers in a variety of fields. Researchers in early intervention have conducted secondary analyses of such data sets to explore relationships between nonmalleable and malleable factors and child outcomes, and to address issues of measurement. Survey data have been used to a lesser extent to examine plausible causal relationships between variables, perhaps due to the increased likelihood of selection bias that results with nonexperimental data. In this article, we use National Early Intervention Longitudinal Study data to demonstrate the use of inverse probability of treatment weighting, a quasi-experimental methodology based on propensity scores that can be used to reduce selection bias and examine plausible causal relationships. We discuss the advantages and disadvantages of this approach, and implications for its use in early intervention research.

Keywords

quasi-experimental methods, propensity score weighting, inverse probability of treatment weighting, early intervention

Introduction

A number of data sets from large-scale, prospective longitudinal studies of young children and families have become available for secondary analysis by researchers in a variety of fields, including early intervention. In addition to the benefit of including data from large samples of children and families who were studied longitudinally, these data sets offer a plethora of variables that quantify nonmalleable and malleable factors potentially related to child or family outcomes (e.g., child and family demographics; pre- and post-natal family and child health histories; child-rearing practices; and child and family experiences during children's early years of development, including public and private services received). Examples of these data sets include the National Early Intervention Longitudinal Study (NEILS; SRI International, 2018), the Pre-Elementary Education Longitudinal Study (PEELS; Institute of Education Sciences, National Center for Special Education Research, n.d.), the Head Start and Early Head Start Family and Child Experiences Survey (Office of Planning, Research & Evaluation, 1997-2018, 2011), and

[1]University of Florida, Gainesville, USA

Corresponding Author:
Crystal D. Bishop, Anita Zucker Center for Excellence in Early Childhood Studies, University of Florida, Gainesville, FL 32611, USA.
Email: crowecd@coe.ufl.edu

the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B; Institute of Education Sciences, National Center for Education Statistics, n.d.-a).

There has been growing support and training for using large-scale data sets from entities such as the Institute of Education Sciences in the Department of Education, the American Educational Research Association, and the Office of Planning, Research, and Evaluation in the Office of Administration for Children and Families in the Department of Health and Human Services. Researchers in the field of early childhood studies, including early intervention, have begun to use large-scale data sets to (a) explore assessment frameworks or measurement issues in early care and education settings (e.g., Bishop, Snyder, & Algina, 2018; Gordon, Fujimoto, Kaestner, Korenman, & Abner, 2013), (b) identify groups of individuals who share similarities in measures of particular variables in specific areas by using person-oriented methodological approaches (e.g., Cook, Roggman, & D'zatko, 2012; McLaughlin, Snyder, & Algina, 2015); and (c) examine correlations between characteristics of children's early care and education experiences and their developmental and learning outcomes (e.g., Burchinal, Peisner-Feinberg, Bryant, & Clifford, 2000; Peisner-Feinberg et al., 2001). To a lesser extent, particularly in early intervention, researchers in early childhood have begun to apply quasi-experimental methods in secondary analyses of large-scale data sets to examine plausible causal relationships between variations in children's early learning experiences and their developmental outcomes (e.g., Ruzek, Burchinal, Farkas, & Duncan, 2014; Sullivan & Field, 2013).

One potential explanation for the underutilization of quasi-experimental methods to explore plausible causal relationships is that data sets used for secondary analyses are often from nonexperimental survey studies, which present challenges related to internal validity threats. Of particular concern is a limitation to draw causal inferences from survey data due to selection bias (Shadish, Cook, & Campbell, 2002). Selection bias is a discrepancy between the estimated treatment effect and the true treatment effect due to systematic pre-intervention differences between members of treated and untreated groups. When participants are randomly assigned to experimental groups, as occurs in experimental studies, pre-intervention differences between groups are assumed to be distributed randomly across groups. When participants are not randomly assigned to groups, observed pre-intervention differences cannot be assumed to be random, and steps must be taken to assess and reduce selection bias.

A number of statistical methodologies exist to reduce selection bias in survey data, allowing researchers to conduct quasi-experimental studies in which treatment effects can be interpreted to draw conclusions about plausible causal relationships between independent and dependent variables. In this article, we focus on the use of propensity score (PS) methods (Rosenbaum & Rubin, 1983) to reduce selection bias. The primary purposes of this article are to provide an introduction to PS methods and to illustrate the relevance of using one particular PS method, inverse probability of treatment weighting, to estimate treatment effects with large-scale survey data. To accomplish these purposes, we provide an overview of treatment effects commonly estimated using survey data, describe sources of selection bias, and illustrate the use of inverse probability of treatment weighting with example survey data.

Rubin's Potential Outcomes Framework for Estimating Treatment Effects

A common way of conceptualizing the issue of estimating treatment effects with survey data is Rubin's Potential Outcomes framework for causal inference (Rubin, 1974). In the following discussion, we use the term *treatment* to refer to the independent variable of interest. Rubin's framework makes the following assumptions: (a) All individuals in the population have potential outcomes for each treatment condition (e.g., treatment vs. no treatment; treatment vs. alternative treatment), (b) the outcomes associated with a particular treatment condition will be observed only in the presence of that condition, and (c) the outcomes associated with no treatment or an

alternative treatment will be observed only in the absence of the treatment (Guo & Fraser, 2014). We describe two treatment effects applicable to early intervention research that are most commonly estimated using Rubin's framework: the average treatment effect (ATE) and the average treatment effect on the treated (ATT).

The ATE is used to estimate the mean difference of potential outcomes for all individuals, given exposure to a particular independent variable. The ATE is estimated using the following formula:

$$ATE = E\left[Y_i^t\right] - E\left[Y_i^c\right], \tag{1}$$

In Equation 1, $E[Y_i^t]$ denotes the expected potential outcome for all individuals if they are exposed to one treatment condition, and $E[Y_i^c]$ is the expected potential outcome for all individuals if they are exposed to the comparison condition (Morgan & Harding, 2006; Winship & Morgan, 1999). For example, a researcher might want to examine the impact of a treatment of attending a voluntary pre-kindergarten program (VPK) for children who are at risk for developmental delays. In this case, the ATE would be the difference between the mean of potential outcomes for *all* children at risk for developmental delays if they attend VPK (i.e., $E[Y_i^t]$ in Equation 1) and the mean of potential outcomes for *all* children at risk for developmental delays if they did not attend VPK (i.e., $E[Y_i^c]$ in Equation 1). However, given children cannot be both participants and nonparticipants in VPK at the same time, only one potential outcome is observed for each child, while the other is missing.

In contrast to ATE, the ATT is used to estimate the mean difference between expected observed outcomes for *individuals who received treatment* and their expected outcomes had they not received treatment. The ATT is estimated using the following formula:

$$ATT = E\left[Y_{iT}^t\right] - E\left[Y_{iT}^c\right], \tag{2}$$

In Equation 2, $E[Y_{iT}^t]$ represents the expected observed outcome for the individuals who received treatment, and $E[Y_{iT}^c]$ is the expected potential outcome for individuals who did not receive treatment. For example, the ATT of receiving preschool special education services for preschool children with disabilities would be the difference between the mean of observed outcomes of children who received preschool special education services (i.e., $E[Y_{iT}^t]$ in Equation 2) and the mean of their potential outcomes if they had not received services (i.e., $E[Y_{iT}^c]$ in Equation 2).

In summary, ATE is used to examine the effect of a treatment for all individuals. In contrast, ATT is used to examine the effect of treatment only for individuals who received the treatment. Either treatment effect can be estimated using survey data; however, the extent to which they can be interpreted with confidence is contingent on the assessment of and adjustment for selection bias.

Selection Bias in Survey Data: Potential Influence on Estimated Treatment Effects

Selection bias occurs when there is systematic variation in baseline characteristics of individuals across levels of the independent variable (Shadish et al., 2002). When selection bias exists, treatment effects are biased, and findings cannot be interpreted with confidence. Two conditions must be met to ensure treatment effects are not afflicted by selection bias: strong ignorability of treatment and the overlap assumption.

Strong Ignorability of Treatment Assignment

Under Rubin's potential outcomes framework, estimates of treatment effect, including ATE and ATT, will be unbiased only if there is strong ignorability of treatment assignment. To achieve

strong ignorability of treatment assignment, the probabilities of selection into treatment conditions (e.g., attend VPK, did not attend VPK) conditional on baseline characteristics of individuals (e.g., parental education, urban or rural residence) must be uncorrelated with the potential outcome distributions (e.g., reading achievement) of treated and untreated individuals (Rosenbaum & Rubin, 1983).

In true experimental studies, randomly assigning individuals to treatment and nontreatment conditions enables differences between baseline characteristics of individuals across conditions to be random rather than systematic (Shadish et al., 2002). In this case, the strong ignorability of treatment assignment is met, allowing the researcher to estimate treatment effects by comparing outcomes of different treatment groups directly. In contrast, data collection for large-scale surveys takes place within naturally occurring mechanisms that are likely to impact individual probabilities of exposure to treatment and individual outcomes in nonrandom ways.

For example, consider a research question focused on examining the impact of family engagement in early intervention programming on children's developmental outcomes. An example of an individual baseline characteristic that might impact exposure to treatment (i.e., family engagement in early intervention programming) and children's developmental outcomes is the extent to which a family believes they can help their child develop and learn. Because it is not possible to randomly assign families to varying levels of family engagement, observed differences in the extent to which families report they can help their child develop and learn cannot be assumed to be random. When this occurs, there is no strong ignorability of treatment, and it is not possible to determine whether observed treatment effects are due to family beliefs regarding their ability to help their child develop and learn or to the independent variable of interest. In such cases, there is no strong ignorability of treatment, and selection bias exists. This issue can become particularly problematic with survey data, where there are numerous variables that might be associated with individual exposure to the independent (treatment) variable of interest and individual outcomes.

The Overlap Assumption

Selection bias is also impacted by violations of the overlap assumption. The overlap assumption requires that every participant has a chance of being in any of the treatment conditions of interest. This implies that the probability of treatment assignment is neither zero nor one for any participant for any treatment condition (Rosenbaum & Rubin, 1983). An example of a violation of this assumption might occur with survey data if a researcher were interested in examining the effect of participating in parent support groups on parental self-efficacy in a sample of parents of children with identified disabilities. If some of the participants in the sample live in rural areas in which parent support groups are not offered, their probability of receiving the treatment is zero, resulting in a violation of the overlap assumption. When such a violation exists, the first assumption under Rubin's framework is violated. As such, it is impossible to achieve strong ignorability of treatment, which results in selection bias.

Addressing Selection Bias in Survey Data Using PS

To obtain unbiased estimates of treatment effect when the above conditions are not met, it is necessary to reduce selection bias by adjusting for covariates (i.e., baseline characteristics) that are related to both an individual's probability of being in a particular treatment group and the outcomes. A traditional approach to remove selection bias in treatment effect estimation due to covariates is to use analysis of covariance (ANCOVA); however, as explained in detail by Schafer and Kang (2008), the ANCOVA approach has several shortcomings. First, the treatment effect does not have a clear interpretation under Rubin's potential outcome as either the ATE or ATT. Second, ANCOVA requires that the form of the relationship between covariates and the

outcome be specified correctly, which is difficult if the number of covariates is large. Third, ANCOVA does not include any built-in checks of whether the overlap assumption is met or covariate distributions were adequately balanced between treated and untreated groups. One alternative to ANCOVA for reducing selection bias in empirical investigation of survey data is the use of PS methods (Ho, Imai, King, & Stuart, 2007). A PS is an individual probability of treatment assignment predicted by observed covariates. In comparison to ANCOVA, PS methods are particularly advantageous when there are a large number of observed covariates that can be used to remove selection bias.

A number of PS methods (e.g., matching, stratification, weighting) have been described in the literature and shown to be efficacious in reducing selection bias in survey data (e.g., Leite, 2016; Stuart, 2010). All PS models can be conceptualized as methods to create observation weights that adjust covariate distributions to be similar across treatment groups and to reduce selection bias (Leite et al., 2015). For a single treatment version, the simplest case is one-to-one matching without replacement, in which treated and matched observations receive a weight of one and unmatched observations receive a weight of zero, and are, therefore, dropped from the analysis. In one-to-many matching, the only difference from one-to-one matching is that matched units receive weights that are the inverse of the number of matches. In PS stratification, the sample is divided into strata based on the PS, and individual weights are created based on the number of treated and untreated individuals in each stratum. Another PS method is inverse probability of treatment weighting, which involves calculating weights for each individual that represent the inverse probability of receiving the treatment they received. These weights are then used as observation weights so that, in the resulting weighted sample, the covariates become unrelated to the treatment status (Leite, 2016). Inverse probability of treatment weighting is used in the illustration that follows.

Reducing Selection Bias Using Inverse Probability of Treatment Weighting

In this article, we describe an application of inverse probability of treatment weighting to increase the utility of using large-scale survey data to examine plausible causal relationships relevant for early intervention. We focus on inverse probability of treatment weighting rather than PS matching or stratification, because inverse probability of treatment weights can be included in the analyses in the same way that survey sample weights are used, making this approach the most compatible with existing software for survey data analysis.

The NEILS (SRI International, 2018) data set is used to illustrate the use of inverse probability of treatment weighting to examine the effect of family engagement in early intervention programming (i.e., treatment) on children's language and literacy status in kindergarten (i.e., outcome). This application was chosen because it highlights a number of important analytical issues relevant to conducting secondary analyses of large-scale survey data to answer questions of plausible causal inference in the field of early intervention. A brief overview of the NEILS study and relevant sources of data for the present illustration are provided, followed by an illustration showing how inverse probability of treatment weighting can be applied to reduce selection bias in the study sample. The intent of this article is not to disseminate research findings but rather to illustrate the use of inverse probability of treatment weighting as a method to reduce selection bias in survey data. We conclude with a discussion of the implications for using this approach in early intervention research.

Introduction to the NEILS Study

The NEILS data set contains data from a nationally representative sample of 3,338 children entering Part C early intervention (birth to 3 years) services from September 1997 through

November 1998 (Javitz, Spiker, Hebbeler, & Wagner, 2002). A three-stage sampling approach was used to select states, counties, and individual children and families for participation in the study. State sampling probability was assigned based on the percentage served of the total children receiving early intervention. States were also sampled to ensure variations in (a) early intervention eligibility criteria, (b) geographic region and population size, (c) early intervention administration agency, and (d) traditionally underrepresented groups receiving early intervention services. Within the sampled states, counties (i.e., local sampling units) were sampled proportional to the county population of children aged birth to 3 years. Within counties, children and families who were receiving early intervention services were sampled at rates that were the inverse of the probability of county selection. Children were recruited for the study at the time they were found eligible for early intervention services. Children eligible to participate in the survey study (a) were less than 31 months of age at the time the first individual family service plan (Individualized Family Service Plan [IFSP]) was signed, (b) had an English- or Spanish-speaking adult in the household who could answer questions about the family and child, and (c) were the only child in the family recruited for the study. The final study sample included 3,338 children and families sampled from 20 states and 93 counties (Javitz et al., 2002).

NEILS data were collected longitudinally from the children's families, early intervention professionals, program directors, and kindergarten teachers (Javitz et al., 2002). Telephone interviews were conducted with the children's families in four waves of data collection: (a) when children entered early intervention, (b) 1 year after entering early intervention, (c) when children were 3 years old, and (d) when children entered kindergarten. At each wave of data collection, families provided information about child and family characteristics, child functioning, participation in early intervention services, and family perspectives about early intervention services. A professional working with the family during the first 6 months of early intervention completed a survey to provide information about himself or herself (e.g., background, training, service delivery). The director of the program serving the NEILS family during this time completed a survey to provide information about the program (e.g., number and type of clients served, service providers employed, location of service provision). The data set also contains information provided by early intervention professionals and kindergarten teachers who provided services to children and families participating in the study. Given the complex sampling design and longitudinal data collection, the NEILS data set contains weights to adjust for sampling and nonresponse rates across data-collection waves.

Illustration Sample and Data

The present illustration began with a subsample of 726 NEILS participants. For purposes of the illustration, only children who had an IFSP for the year they turned 3 years old and whose families reported they continued receiving preschool special education and related services after turning 3 were included. Data for the illustration were drawn from (a) family interviews when the children entered early intervention, (b) family interviews when the children were 3 years old, (c) family interviews when the children entered kindergarten, and (d) surveys completed by the children's kindergarten teachers. These sources were chosen because they contained interview or self-report items that represented (a) variables related to family involvement in early intervention programming (the independent/treatment variable), (b) variables related to children's language and literacy skills in kindergarten (the outcome variable), or (c) variables hypothesized as confounders likely to have an impact on the independent variable, the outcome variable, or both the independent variable and the outcome variable. Items pertaining to these categories were used to (a) define the independent variable, (b) estimate PSs to reduce selection bias, or (c) generate a single outcome variable for use in the estimation of treatment effect.

Independent (Treatment) Variable

The independent variable of interest ("treatment") was family engagement in the development of the child's IFSP and transition plan. Rather than a binary indicator of engagement, the level of family engagement was measured by five questions in the family interview when children turned 3 years old (NEILS, 2000). Each question had three response categories. We defined the treatment variable as a categorical latent variable, and used latent class analysis (LCA) to determine categories of family involvement in the development of the child's IFSP and transition plan. We determined three classes of family engagement: (a) families collaborating with early intervention professionals to determine early intervention programming ($n = 324$); (b) professionals determining the early intervention programming, with limited participation from the family ($n = 308$); and (c) families collaborating with professionals to determine early intervention programming, but taking a more active role in decisions regarding services and supports the child would receive after transitioning from early intervention ($n = 94$). Additional information about the LCA and the classes of family engagement is available from the first author upon request.

For this illustration, the 632 families from the first two classes are used to represent alternative treatment groups, each describing who made decisions about early intervention programming. The third group was not included in the present illustration due to violations of the overlap assumption, which were associated with the imbalance in the size of this group compared to the other two. In the illustrated analyses, we use *professionals* to refer to the group in which professionals determined early intervention programming and *collaborative* to refer to the group in which families and professionals collaborated to determine early intervention programming.

Dependent Variable

For purposes of illustration, we used data from the Kindergarten Teacher Survey (NEILS, 2002) to compute a single outcome variable representing children's language and literacy status in kindergarten. The Kindergarten Teacher Interview includes a nine-item scale on which teachers ranked children's language and literacy skills, knowledge, and behaviors. Teachers ranked child proficiency on each item as 1 (*not yet demonstrating the skill)*, 2 (*beginning to demonstrate the skill*), 3 (*demonstration in progress*), 4 (*intermediate demonstration*), or 5 (*proficient*). Teachers could also respond that an item was not applicable. These items are part of a larger scale called the Academic Rating Scale (ARS; National Center for Education Statistics, n.d.), which was developed for use in the Early Childhood Longitudinal Study–Kindergarten Cohort (ECLS-K; Institute of Education Sciences, National Center for Education Statistics, n.d.-b). Item-level scores were averaged across all items for which the teacher provided a ranking of the child's proficiency, which resulted in a single, continuous outcome variable.

Application of Inverse Probability of Treatment Weighting to Reduce Selection Bias

The steps of applying inverse probability of treatment weighting to remove selection bias are (a) select covariates for the PS model, (b) estimate PSs, (c) check the overlap assumption and examine the overlap of PSs across groups, (d) calculate inverse probability of treatment weights, (e) check for extreme weights, and (f) assess balance of covariates across the two treatment groups. Each of these steps is described below. Given the illustrative focus of the present article, we have not included a discussion of some important design issues related to conducting quasi-experimental studies (e.g., stable unit treatment value assumption, sampling weights, controlling for other internal validity threats). Steps a through f were implemented using the R statistical software (R Core Development Team, 2017). The R code used was similar to that of Leite (2016) and is available from the first author. Alternatively, the same PS analyses could have been implemented with other major statistical packages such as SAS 9.3, SPSS Statistics Version 25, or Stata Release 15.

Selecting Covariates for the PS Model

The efficacy of PS methods to reduce selection bias is dependent, in large part, on the inclusion of the appropriate observed baseline covariates in the PS model. Guidance on which variables should be included in the PS model varies in the applied literature. In general, three types of variables might be considered: (a) all covariates hypothesized to be associated with selection into treatment, (b) all covariates hypothesized to be associated with the outcome (i.e., potential confounders), and (c) all covariates hypothesized to be associated with both the selection into treatment and the outcome (i.e., true confounders; Austin, 2011).

Only true confounders produce selection bias in treatment effect estimates if omitted, so identifying and including them in the PS model is important. Some researchers have noted that when potential or true confounders are left out of the model, overlap of PSs is lessened, and when both potential and true confounders are included, the precision of estimated treatment effects increases (Austin, Grootendorst, & Anderson, 2007). Brookhart and colleagues (2006) suggested potential confounders should always be included in the PS model. Austin (2011) noted it is difficult to accurately separate true confounders, potential confounders, and covariates that affect only treatment exposure, but that most subject-level baseline covariates are likely to affect both treatment exposure and treatment outcomes. Following this perspective, some researchers have recommended including all baseline covariates in the PS model, particularly given the goal is to estimate probabilities of treatment assignment rather than interpreting the effects of covariates (Austin, 2011; Schafer & Kang, 2008).

The baseline covariates for the PS model in the present example were drawn from family interviews conducted when the children were enrolled in early intervention, which includes 1,043 variables. Given the extensive number of variables, we selected variables that were most likely to be either potential or true confounders. We identified 44 variables from the Family Enrollment Interview (see Table 1; NEILS, 2000) that, based on previous research and theory, were potential sources of selection bias. The covariates in this PS model included categorical and continuous variables related to maternal education, maternal employment, maternal age at the time of the child's birth, family income, family social supports, parental self-efficacy related to supporting the child's development, family expectations for the child, and the child's acquisition of developmental milestones at entry into early intervention (see Table 1).

Estimating PSs

Estimating PSs involves modeling the selection bias mechanism to calculate each individual's probability of receiving treatment, conditional on the observed baseline covariates included in the model. For a single treatment in which the treatment indicator is a dichotomous variable indicating treated or untreated status, PSs can be estimated with parametric models such as logistic regression and probit regression, or nonparametric data-mining methods such as regression trees, random forests, and generalized boosted modeling (McCaffrey, Ridgeway, & Morral, 2004).

In the present illustration, the treatment variable is dichotomous; thus, we chose to use logistic regression to estimate PSs. We used a linear model with no interactions:

$$\operatorname{logit}\left(Z_i = 1 \mid X\right) = \beta_0 + \sum_{1}^{k} \beta_k X_{ki}, \tag{3}$$

where Z_i indicates the treatment group membership, β_0 is the intercept, and β_k is the regression coefficient of each observed covariate X_{ki}.

The PS $e_i(X)$ is obtained with:

$$e_i\left(X\right) = \frac{\exp\left(\operatorname{logit}\left(Z_i = 1 \mid X\right)\right)}{1 + \exp\left(\operatorname{logit}\left(Z_i = 1 \mid X\right)\right)}. \tag{4}$$

Table 1. Covariate Imbalance Before and After Inverse Probability of Treatment Weighting.

Covariate	No. of categories[a]	Imbalance before weighting[b] (*d*)	Imbalance after weighting[c] (*d*)
Education level of primary female caregiver	7	4 (–.21 to .08)	0 (–.04 to .02)
Age of primary female caregiver at birth	—	Balanced (.02)	Balanced (.03)
Hours worked per week by primary female caregiver	4	0 (–.02 to .03)	0 (–.02 to .03)
Primary female caregiver taking any courses	2	0 (–.02 to .02)	0 (–.01 to .01)
Primary female caregiver's marital status	4	3 (–.14 to .13)	0 (–.03 to .04)
Family's overall life situation now	5	2 (–.10 to .11)	0 (–.01 to .01)
Child's overall life situation now	7	2 (–.03 to .11)	1 (–.01 to .01)
Expectation for family's future life situation	7	3 (–.08 to .11)	2 (–.05 to .06)
Expectations for child's future life situation	7	3 (–.09 to .10)	0 (–.03 to .03)
Know how to care for child's basic needs	4	1 (–.03 to .06)	0 (–.01 to .04)
Know how to help child learn and develop	5	3 (–.08 to .10)	0 (–.05 to .04)
Know how to work with professionals and advocate	5	3 (–.10 to .10)	2 (–.04 to .07)
Have relatives or friends for support	5	3 (–.11 to .11)	1 (–.04 to .07)
Difficulty finding what to do about child's behavior	5	2 (–.06 to .13)	0 (–.02 to .02)
Know what to do if child is not getting services	5	2 (–.10 to .10)	0 (–.04 to .02)
Have relatives, friends, or someone else who helps me deal	5	2 (.06 to .07)	0 (–.04 to .04)
Number of people living in household	—	Imbalanced (.07)	Balanced (–.01)
Receive Aid to Families with Dependent Children, Temporary Assistance for Needy Families, or welfare now or in the past year	3	2 (–.05 to .06)	2 (–.06 to .05)
Receiving food stamps now	2	0 (<.01)	2 (–.05 to .05)
Number of adults living in household	—	Imbalanced (–.07)	Balanced (.01)
Get food vouchers from WIC now	3	3 (–.08 to .07)	0 (–.03 to .04)
Receive money for child from Supplemental Security Income now	3	1 (–.05[d] to .06)	0 (–.03 to .04)
Way current transportation meets needs	4	2 (–.10 to .11)	2 (–.08 to .06)
Household income	—	Balanced (.04)	Balanced (.01)
Number of children living in household	—	Imbalanced (.11)	Balanced (–.02)
Number of other children with special needs	—	Balanced (.04)	Balanced (–.05[d])
Total hours in child care at enrollment	—	Imbalanced (.10)	Balanced (.03)
Babbles	3	2 (–.20 to .09)	0 (–.03 to .03)
Holds up toys or objects to show	3	3 (–.20 to .13)	0 (–.02 to .02)
Uses motions or gestures to communicate	3	2 (–.20 to .20)	0 (–.01 to .02)
Says "mama" or "dada"	3	2 (–.18 to .15)	0 (–.03 to .02)
Repeats or imitates a word	3	2 (–.14 to .13)	1 (–.05 to .02)
Says five or more words other than "mama" or "dada"	3	2 (–.10 to .12)	0 (–.04 to .03)
Asks "what's that" questions	3	3 (–.14 to .11)	1 (–.06 to .04)
Says at least 20 different words	3	2 (–.10 to .09)	0 (–.03 to .03)
Uses any pronouns	3	2 (–.13 to .16)	1 (–.05 to .04)
Says two or three words in a sentence	3	2 (–.05 to .09)	1 (–.09 to .04)
Looks at something you hold	3	3 (–.14 to .17)	0 (–.03 to .02)
Looks up or smiles when you say name	3	3 (–.24 to .24)	0 (–.03 to .02)
Looks at things you point to	3	3 (–.11 to .19)	0 (–.03 to .02)
Responds to simple gestures	3	2 (–.20 to .21)	0 (–.01 to .02)
Points to things you name	3	2 (–.21 to .21)	0 (–.04 to .03)
Responds to simple verbal request	3	2 (–.21 to .02)	0 (–.03 to .03)
Follows a two-step verbal direction	3	2 (–.20 to .20)	0 (–.04 to .02)
Total	143	84	16

Note. Baseline covariates drawn from the NEILS Family Enrollment Interview (NEILS, 2000). NEILS = National Early Intervention Longitudinal Study.

[a]For categorical variables, numeric values represent the number of response categories; continuous variables are represented by "—."

[b]For categorical variables, values represent the number of response categories that were imbalanced before inverse probability of treatment weighting, followed by the range of standardized effect sizes across response categories. Continuous variables are noted as balanced or imbalanced, followed by the standardized effect size.

[c]For categorical variables, values represent the number of response categories that remained imbalanced after inverse probability of treatment weighting, followed by the range of standardized effect sizes across response categories. Continuous variables are noted as balanced or imbalanced, followed by the standardized effect size.

[d]Actual value rounds up to .05.

Given the large number of covariates included in PS models, we recommend an initial model that includes only main effects and then assessing overlap and balance of covariates to determine whether modifications to the PS model are necessary (i.e., addition of interactions or polynomial terms). Following recommendations in Leite (2016), missing data on the covariates was handled using multiple imputation, in which the logistic regression model was estimated separately for each of 10 imputed data sets, and the mean of PSs across data sets was taken.

Checking the Overlap Assumption and Examining Common Support

After PSs have been estimated, they must be examined to ensure the overlap assumption (Rosenbaum & Rubin, 1983) is not violated. In order for this assumption to be met, the distribution of PSs for each group must be between zero and one, with no cases having an estimated PS of either zero or one. In the present example, PSs ranged from 6.4×10^{-8} to .99999998, and thus, the overlap assumption is met. It is important to note, however, that the PSs for some individuals were very near zero or one, resulting in extremely high probability for being assigned to one group over the other. Although technically the overlap assumption is met, the fact that some cases have high PSs might result in a reduced area of common support across groups. The area of common support is the area of the distribution of PSs in which values exist for both treatment groups. Assessing the area of common support is an essential step when using PS methods, because generalizations about the treatment effect can be made only within the area of common support. If the area of common support is small, ignoring this problem leads to biased treatment effect estimates; however, removing observations outside of the area of common support reduces the generalizability of the findings (Crump, Hotz, Imbens, & Mitnik, 2009).

Figure 1 illustrates the area of common support of the PS distributions for our illustration. PSs for the collaborative group ranged from 6.4×10^{-8} to .91. PSs for the professionals group ranged from .13 to .99999998 and thus, the area of common support in this example is the sample of individuals for whom PSs ranged from .13 to .91. This included 599 (94.8%) individuals out of the 632 originally included in the sample. Seventeen individuals in the collaborative group had PSs lower than .13, and 16 individuals assigned to the professionals group had PSs greater than .91. For PS methodologies that rely on matching individuals with similar PS across groups, all cases outside the area of common support would necessarily be dropped from the analysis. The inverse probability of treatment weighting approach allows the possibility of maintaining the original sample, because it does not require matching of PS across groups; however, it is not immune from problems arising as a result of limited overlap of PS across groups.

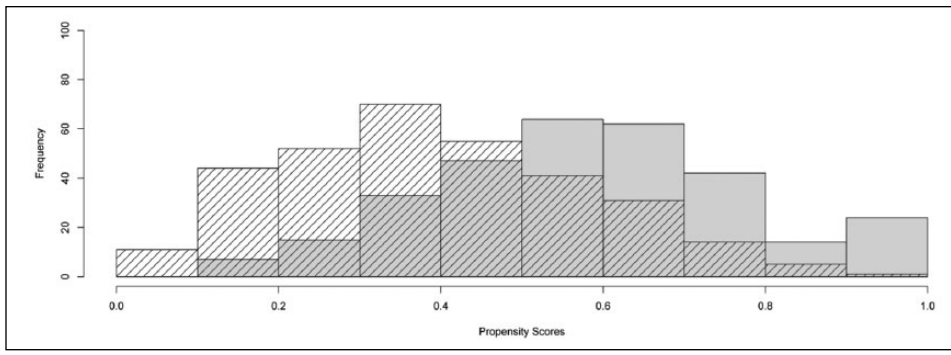

Figure 1. Histogram of overlap between collaborative (patterned) and professional (gray) groups.

Calculating and Assessing Inverse Probability of Treatment Weights

Inverse probability of treatment weights are calculated for each individual in a sample. Each person's weight is equal to the inverse of the probability of receiving the treatment he or she actually received (Leite, 2016), and is derived using the following formula:

$$w_i = \frac{Z_i}{\hat{e}_i(\mathrm{X})} + \frac{1 - Z_i}{1 - \hat{e}_i(X)}, \tag{5}$$

where w_i is the individual weight; Z_i is the treatment indicator, which takes values of 0 or 1; and $\hat{e}_i(X)$ is the estimated PS.

A potential problem with the inverse probability of treatment weighting approach is the presence of extreme weights, which result in inflated standard errors and may also increase bias (Harder, Stuart, & Anthony, 2010). An extreme weight occurs when a treated individual has a very low estimated probability of being treated, or an untreated individual has a very high probability of being treated.

Extreme weights may be due to misspecification or the PS model (Lee, Lessler, & Stuart, 2011) or poor common support. For this reason, it is important to analyze the distribution of weights in the sample to determine the extent to which extreme weights are present. There is no precise definition of which values of weights can be considered extreme, and this determination is dependent on sample size (e.g., a weight of 100 may be extreme if the sample size is 200, but not if it is 200,000). Some solutions have been proposed to deal with extreme weights, which can also be used to determine whether the results are sensitive to these weights. These include respecifying the PS model, truncating weights at a certain percentile (Lee at al., 2011), and trimming the sample outside of the [0.1, 0.9] interval of the PS (Crump et al., 2009).

In the present illustration, no extreme weights were identified. Weights for the entire illustrative sample ranged from 1.00 to 11.88. In the group in which early intervention programming was determined collaboratively by the family and early intervention professionals, weights ranged from 1.00 to 11.88. In the group in which early intervention programming was determined primarily by early intervention professionals, weights ranged from 1.00 to 9.272.

Checking Covariate Balance

The effectiveness of PS methods for reducing selection bias is evaluated by examining the balance of covariates across groups after the method has been applied (Stuart, 2010). If covariates are not balanced across treatment conditions after applying a PS method, then the PS model must be revised by adding interactions between covariates and nonlinear effects of covariates. If logistic regression does not lead to adequate covariate balance, estimation of PSs with data-mining methods may be useful (McCaffrey et al., 2004).

Approaches used to assess covariate balance include graphics, descriptive statistics, inferential statistics, and graphic diagnostic. The approach most commonly used for covariate balance diagnostic is the standardized mean difference of treated and untreated groups on each covariate, which is a descriptive statistic. Inferential statistics, such as a *t* test of the difference between means, are not recommended because these tests refer to a population, and covariate balance is a property of the sample. Furthermore, inferential statistics are affected by sample size, which is not desirable when evaluating covariate balance. For continuous covariates, graphic diagnostic of covariate balance consists of examining empirical QQ-plots of the quantiles of the treated group versus the quantiles of the untreated group. Points on a 45-degree line indicate adequate covariate balance. For categorical covariates, overlapping bar plots of categories of treated and untreated groups can be used.

For the present illustration, we assessed balance by calculating standardized effect sizes for the weighted differences in means or proportions across the treatment conditions for all of the 44 covariates shown in Table 1. Standardized effect sizes were calculated with and without the use of inverse probability of treatment weighting. The mean differences were standardized with the standard deviation of the professional group, but the pooled standard deviation could also have been used (Leite, 2016). For categorical variables, we evaluated covariate balance in proportions for each category. Altogether, the continuous covariates and categories of categorical covariates resulted in 143 standardized effect sizes. We applied the standards for quasi-experimental designs of the What Works Clearinghouse (U. S. Department of Education, Institute of Education Sciences, & What Works Clearinghouse, 2013) to evaluate covariate balance. A covariate is considered balanced across treated groups without any need for further adjustment if the absolute value of the standardized effect size is less than 0.05 standard deviations. For covariates with standardized effect sizes between 0.05 and 0.25 standard deviations, balance is considered adequate if the covariate is also included in the outcome model.

Table 1 illustrates covariate balance before and after the application of inverse probability of treatment weighting. For categorical covariates, the number of categories with standardized effect sizes with absolute values greater than 0.05 before and after weighting are shown. Continuous covariates are shown as either "balanced" or "unbalanced." Standardized effect sizes decreased following application of inverse probability of treatment weighting, indicating the covariate distributions became more similar across groups when the observations were weighted with the inverse probability of treatment weights. Without inverse probability of treatment weighting, the maximum standardized effect size was 0.24, and 84 out of 144 covariates/categories had effect sizes with absolute values above 0.05. After observations were weighted with inverse probability of treatment weights, only 16 categories from 11 categorical covariates had effect sizes with absolute values above 0.05. The maximum standardized effect size of these categories was 0.08, which is within the range of effect sizes in which balance is considered adequate, provided they are included in the outcome model.

Estimation of Treatment Effects

After examining covariate balance, we proceeded to estimate the effect of family involvement in early intervention programming on children's language and literacy status. To illustrate the varying levels of selection bias removal, we estimated the treatment effect using three different models. The first model was $y_i = \beta_0 + \beta_1 Z_i + \varepsilon_i$, where β_0 is the mean of the collaborative group, and β_1 is the difference between the mean of the collaborative group and the professional group, which is the treatment effect. The second model was specified as in the first model, but it is a weighted regression model in which inverse probability of treatment weighting removed selection bias. In the second model, β_0 is the weighted mean of the collaborative group, and β_1 is the weighted difference between the collaborative and professional group. Following the recommendations articulated in the What Works Clearinghouse (U.S. Department of Education, Institute of Education Sciences, & What Works Clearinghouse, 2013), the third model used inverse probability of treatment weights and also included the covariates that were not balanced by weighting: $y_i = \beta_0 + \beta_1 Z_i + \gamma_1 X_{1i} + \cdots \gamma_c X_{ci} + \varepsilon_i$, where $X_{1i} \ldots X_{ci}$ are covariates. This latter model is said to be doubly robust, because it includes two mechanisms (i.e., inverse probability of treatment weighting, controlling for unbalanced covariates) for removing selection bias (Kang & Schafer, 2007). Doubly robust methods consistently estimate the treatment effect if either the PS model or the outcome model are correctly specified (Robins & Rotnitzky, 1995).

Results from the three models are shown in Table 2. For our example, the findings from the three different models are very similar, with each model resulting in a statistically significant

Table 2. Logistic Regression Model Results With and Without Inverse Probability of Treatment Weighting.

Model	Intercept (collaborative mean)	Effect of professionals (mean difference)	Standardized mean difference	*SE*
1. No adjustment	2.409	0.226	0.187	.099
2. IPTW	2.395	0.218	0.180	.106
3. IPTW plus covariates	2.420	0.222	0.200	.096

Note. All effects are statistically significant at $\alpha = .05$. IPTW = Inverse Probability of Treatment Weighting.

effect of professionals making program decisions on child outcomes. Although the three models yielded nearly equal results, they cannot be interpreted with equal confidence. Model 3 yields the most robust findings with respect to producing outcomes that are not impacted by selection bias, because it includes two controls for this validity threat.

As illustrated in Table 1, the balance of baseline covariates improved considerably with the application of inverse probability of treatment weighting, reducing the threat of selection bias. In the present example, the impact of the unbalanced covariates on the estimates of treatment effect appears to be minimal; however, this is not always the case with quasi-experimental research. It is incumbent on the researcher to make adjustments like those presented in this example to ensure the largest removal of selection bias and strengthening of internal validity of estimated treatment effects.

Discussion

In this example, we illustrated procedures for generating PSs and applying inverse probability of treatment weighting and discussed the utility of these approaches for reducing selection bias and evaluating treatment effects using large-scale survey data. Additional considerations for these illustrated approaches are discussed, in addition to the relevance of these approaches for early intervention research.

Considerations for Applying PS and Inverse Probability of Treatment Weighting

Although PS methods in general, and inverse probability of treatment weighting in particular, offer advantages for reducing selection bias in survey data, some limitations must be considered. First, this approach assumes there are no unmeasured true confounders (Rosenbaum & Rubin, 1983). In practice, it is unlikely this assumption will be met. A sensitivity analysis method such as the one proposed by Carnegie, Harada, and Hill (2016) and illustrated with inverse probability of treatment weighting by Leite (2016) can be used to determine whether conclusions would change if true confounders of different strength were omitted. Second, if the measurement of covariates is unreliable, including them in a PS model will fail to remove bias; however, use of latent variable models for covariates can ameliorate this problem (Leite, 2016). A third consideration when using PS methods is the impact of the PS model specification. The extent to which covariate balance is achieved depends on the accuracy of the PS model. As such, it is recommended that multiple methods to estimate PS (e.g., logistic regression, data mining) are compared with respect to covariate balance. A final consideration for using the illustrated approaches is that they are aimed only at reducing selection bias. To ensure confidence in interpretability of effects with respect to plausible causal relationships with nonexperimental data, it is incumbent on the researcher to consider and control for other threats to the validity of findings.

Relevance for Early Intervention Research

Despite the noted limitations of the PS methods, the application of these methods in quasi-experimental studies holds promise for expanding early intervention research. The availability of large-scale survey data with variables of interest to researchers in early intervention offers opportunities to conduct exploratory evaluations of the differential effects of treatment programs used by young children with or at risk for developmental delays and their families. One particular advantage of these data sets is that they offer much larger sample sizes and number of variables than are typically feasible to obtain in randomized controlled trials. As such, they have the potential to provide information about how manipulation of malleable factors might influence outcomes for children and families. This information can be used subsequently to develop new interventions or to adapt existing interventions and to inform the conduct of single-case or group experimental designs.

Use of inverse probability of treatment weighting is particularly advantageous because of its versatility. In addition to the application described and illustrated in the present article, inverse probability of treatment weighting can be used with more complexity and in combination with other statistical methodologies to control for additional validity threats. For example, inverse probability of treatment weights can be multiplied by sampling weights to allow sample estimates to generalize to a national population and to account for potential violations of independence due to clustering of individuals within primary sampling units. The inverse probability of the treatment weighting approach can also be used with more complex modeling approaches that might be used to examine issues of interest to early intervention researchers (e.g., structural equation modeling, multilevel modeling; Leite, 2015; Leite et al., 2015; Thoemmes & West, 2011). Finally, use of inverse probability of treatment weighting is not limited to examining dichotomous, single-application treatments. It can be applied to examine the impact of multiple doses of treatment or intervention or of exposure interventions that vary over time (Robins, Hernan, & Brumback, 2000), as is common in early intervention. These applications might be particularly useful to inform the development of adaptive interventions, which are altered systematically and repeatedly over time, based on the changing needs of the intervention recipient (Nahum-Shani et al., 2012).

Conclusion

The increasing availability of comprehensive technical guidance and illustrations of quasi-experimental methodologies with a variety of statistical software (e.g., Faries, Leon, Haro, & Obenchain, 2010, with SAS; Guo & Fraser, 2014, with Stata; Leite, 2016, with R) provides a mechanism for examining plausible causal relationships within large-scale data sets. Although these methods are not without limitation, applications like those described in the present article can be conducted with rigor to illuminate future avenues of research and development or to answer important questions regarding the plausible treatment effects of current programs and interventions available to young children with or at risk for developmental delays and their families.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The work reported in this article was supported, in part, by a postdoctoral research fellowship training grant to the University of Florida (R324B120002) from the Institute of Education Sciences. The opinions expressed are those of the authors, not the funding agency, and no official endorsement should be inferred.

References

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*, 399-424. doi:10.1080/00273171.2011.568786

Austin, P. C., Grootendorst, P., & Anderson, G. M. (2007). A comparison of the ability of different models to balance measured variables between treated and untreated subjects: A Monte Carlo study. *Statistics in Medicine*, *26*, 734-753. doi:10.1002/sim.2580

Bishop, C., Snyder, P. A., & Algina, J. (2018). *Exploring measurement invariance for the ECERS across two types of preschool classrooms*. Manuscript submitted for publication.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Sturmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, *163*, 1149-1156. doi:10.1093/aje/kwj149

Burchinal, M. R., Peisner-Feinberg, E., Bryant, D. M., & Clifford, R. (2000). Children's social and cognitive development and child-care quality: Testing for differential associations related to poverty, gender, or ethnicity. *Applied Developmental Science*, *4*, 149-165. doi:10.1207/S1532480XADS0403_4

Carnegie, N. B., Harada, M., & Hill, J. L. (2016). Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *Journal of Research on Educational Effectiveness*, *9*, 395-420.

Cook, G. A., Roggman, L. A., & D'zatko, K. (2012). A person-oriented approach to understanding dimensions of parenting in low-income mothers. *Early Childhood Research Quarterly*, *27*, 582-595.

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, *96*, 187-199. doi:10.1093/biomet/asn055

Faries, D. E., Leon, A. C., Haro, J. M., & Obenchain, R. (2010). *Analysis of observational health care data using SAS*. Cary, NC: SAS Institute.

Gordon, R. A., Fujimoto, K., Kaestner, R., Korenman, S., & Abner, K. (2013). An assessment of the validity of the ECERS–R with implications for measures of child care quality and relations to child development. *Developmental Psychology*, *49*, 146-160. doi:10.1037/a0027899

Guo, S., & Fraser, M. W. (2014). *Propensity score analysis: Statistical methods and applications* (2nd ed.). Thousand Oaks, CA: Sage.

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*, *15*, 234-249.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*, 199-236.

Institute of Education Sciences, National Center for Education Statistics. (n.d.-a). *Early childhood longitudinal program: Birth cohort (ECLS-B)*. Retrieved from http://nces.ed.gov/ecls/birth.asp

Institute of Education Sciences, National Center for Education Statistics. (n.d.-b). *Early childhood longitudinal program: Kindergarten class of 1998-1999 (ECLS-K)*. Retrieved from http://nces.ed.gov/ecls/kindergarten.asp

Institute of Education Sciences, National Center for Special Education Research. (n.d.). *Pre-Elementary Education Longitudinal Study*. Retrieved from https://ies.ed.gov/ncser/projects/datasets_peels.asp

Javitz, H., Spiker, D., Hebbeler, K., & Wagner, M. (2002). *National Early Intervention Longitudinal Study sampling and weighting procedures: Enrollment form, family interview, service records* (NEILS Methodology Report No. 1). Menlo Park, CA: SRI International.

Kang, J., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, *22*, 523-539. doi:10.1214/07-STS227

Lee, B. K., Lessler, J., & Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PLoS ONE*, *6*(3), e18174.

Leite, W. L. (2015). Latent growth modeling of longitudinal data with propensity score matched groups. In W. Pan & H. Bai (Eds.), *Propensity score analysis: Fundamentals, developments, and extensions* (pp. 191-216). New York, NY: Guilford.

Leite, W. L. (2016). *Practical propensity score methods using R*. Thousand Oaks, CA: Sage.

Leite, W. L., Jimenez, F., Kaya, Y., Stapleton, L. M., MacInnes, J. W., & Sandbach, R. (2015). An evaluation of weighting methods based on propensity scores to reduce selection bias in multilevel observational studies. *Multivariate Behavioral Research*, *50*, 265-284. doi:10.1080/00273171.2014.991018

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, *9*, 403-425.

McLaughlin, T. W., Snyder, P. A., & Algina, J. (2015). Characterizing early childhood disabilities in a nationally representative sample using functional profiles. *Exceptional Children*, *81*, 471-488. doi:10.1177/0014402914563696

Morgan, S. L., & Harding, D. J. (2006). Matching estimators of causal effects: Prospects and pitfalls in theory and practices. *Sociological Methods & Research*, *35*, 3-60. doi:10.1177/0049124106289164

Nahum-Shani, I., Qian, M., Almirall, D., Pelham, W. E., Gnagy, B., Fabiano, G., . . .Murphy, S. (2012). Q-learning: A data analysis method for constructing adaptive interventions. *Psychological Methods*, *17*, 478-494. doi:10.1037/a0029373

National Center for Education Statistics. (n.d.). *Kindergarten Teacher Questionnaire* (Fall) (OMB No. 1850-0719). Retrieved from https://nces.ed.gov/ecls/kinderassessments.asp

National Early Intervention Longitudinal Study. (2000). *Family transition interview* (Telephone interview and manual specifications). Retrieved from https://www.sri.com/work/publications/national-early-intervention-longitudinal-study-neils-family-transition-interview

National Early Intervention Longitudinal Study. (2002). *Kindergarten teacher survey section A: General and child progress* (Survey instrument). Menlo Park, CA: SRI International.

Office of Planning, Research & Evaluation. (1997-2018). *Head Start Family and Child Experiences Survey (FACES), 1997-2018: Project overview*. Retrieved from https://www.acf.hhs.gov/opre/research/project/head-start-family-and-child-experiences-survey-faces

Office of Planning, Research & Evaluation. (2011). *Baby FACES 2009: Learning as we go: A first snapshot of Early Head Start programs, staff, families, and children, Volume I: First report* (OPRE 2011-7). Retrieved from https://www.acf.hhs.gov/opre/resource/baby-faces-2009-learning-as-we-go-a-first-snapshot-of-early-head-start

Peisner-Feinberg, E. S., Burchinal, M. R., Clifford, R. M., Culkin, M. L., Howes, C., Kagan, S. L., & Yazejian, N. (2001). The relation of preschool child-care quality to children's cognitive and social developmental trajectories through second grade. *Child Development*, *72*, 1534-1553.

R Core Development Team. (2017). *A language environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available from http://www.R-project.org

Robins, J. M., Hernan, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11*, 550-560.

Robins, J. M., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, *90*, 122-129.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41-55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, *66*, 688-701.

Ruzek, E., Burchinal, M., Farkas, G., & Duncan, G. J. (2014). The quality of toddler child care and cognitive skills at 24 months: Propensity score analysis results from the ECLS-B. *Early Childhood Research Quarterly*, *29*, 12-21.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, *13*, 279-313. doi:10.1037/a0014268

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

SRI International. (2018). *National Early Intervention Longitudinal Study (NEILS)*. Retrieved from http://www.sri.com/work/projects/national-early-intervention-longitudinal-study-neils

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*, 1-21.

Sullivan, A. L., & Field, S. (2013). Do preschool special education services make a difference in kindergarten reading and mathematics skills? A propensity score weighting analysis. *Journal of School Psychology*, *51*, 243-260. doi:10/1016/j.jsp.2012.12.004

Thoemmes, F. J., & West, S. G. (2011). The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research*, *46*, 514-543. doi:10.1080/00273171.2011.569395

U.S. Department of Education, Institute of Education Sciences, & What Works Clearinghouse. (2013). *What works clearinghouse: Procedures and standards handbook* (Version 3.0). Washington, DC: author.

Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology*, *25*, 659-706.