

Using Propensity Score Analysis of Survey Data to Estimate Population Average Treatment Effects: A Case Study Comparing Different Methods

Evaluation Review
2020, Vol. 44(1) 84-108
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0193841X20938497
journals.sagepub.com/home/erx



Nianbo Dong¹ , Elizabeth A. Stuart^{2,3,4},
David Lenis⁵, and Trang Quynh Nguyen^{2,3}

Abstract

Background: Many studies in psychological and educational research aim to estimate population average treatment effects (PATE) using data from

¹ School of Education, University of North Carolina at Chapel Hill, NC, USA

² Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

³ Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

⁴ Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

⁵ Flatiron Health, New York, NY, USA

Corresponding Author:

Nianbo Dong, School of Education, University of North Carolina at Chapel Hill, 116 Peabody Hall, CB 3500, Chapel Hill, NC 27599, USA.

Email: dong.nianbo@gmail.com

large complex survey samples, and many of these studies use propensity score methods. Recent advances have investigated how to incorporate survey weights with propensity score methods. However, to this point, that work had not been well summarized, and it was not clear how much difference the different PATE estimation methods would make empirically.

Purpose: The purpose of this study is to systematically summarize the appropriate use of survey weights in propensity score analysis of complex survey data and use a case study to empirically compare the PATE estimates using multiple analysis methods that include ordinary least squares regression, weighted least squares regression, and various propensity score applications. **Methods:** We first summarize various propensity score methods that handle survey weights. We then demonstrate the performance of various analysis methods using a nationally representative data set, the Early Childhood Longitudinal Study–Kindergarten to estimate the effects of preschool on children’s academic achievement. The correspondence of the results was evaluated using multiple criteria. **Results and Conclusions:** It is important for researchers to think carefully about their estimand of interest and use methods appropriate for that estimand. If interest is in drawing inferences to the survey target population, it is important to take the survey weights into account, particularly in the outcome analysis stage for estimating the PATE. The case study shows, however, not much difference among various analysis methods in one applied example.

Keywords

equivalence test, population average treatment effects, propensity scores, survey weights, complex surveys

Large-scale, complex survey designs have been widely used in psychological and educational research. This includes research on early childhood education as well as on higher education, and domestic and cross-country education research. Commonly analyzed examples of such surveys include the Early Childhood Longitudinal Study (ECLS), the National Education Longitudinal Study of 1988, High School and Beyond, the Trends in International Mathematics and Science Study, the Progress in International Reading Literacy Study, and the Program for International Student Assessment. These surveys aim to collect information on a sample that is formally representative of its target population, and they do this using a sampling frame and sample design. In large-scale surveys, the sample design may be

complicated, perhaps including a multistage sampling design (e.g., stratified clustered sampling, with rural, suburban and urban schools sampled separately and then students selected within schools), and perhaps with some groups oversampled to ensure large enough samples in those groups. The resulting individual survey weights reflect these varying probabilities of selection and adjust for nonresponse. Those survey design elements then need to be accounted for in analyses in order to draw inferences relevant to the target population (see, e.g., Hansen et al., 1983; Holt et al., 1980; Kish & Frankel, 1974; Korn & Graubard, 1995a, 1995b; Scott & Holt, 1982).

In addition to descriptive and correlational analyses, researchers often want to draw causal inferences using survey data; that is, they want to use survey data to be able to estimate the effects of nonrandomized interventions, risk factors, or exposures. Propensity score methods (Rosenbaum & Rubin, 1983) are a key tool often used to help estimate causal effects in nonrandomized studies. The propensity score is the probability of treatment assignment conditional on observed baseline covariates. Conditional on the propensity score, the distribution of observed baseline covariates will be similar between treated and untreated subjects (Rosenbaum & Rubin, 1983). One way to think about the complication of estimating population treatment effects using data from a complex survey is that when there are heterogeneities in treatment assignment and treatment effects in the sampled population, failure to take into account survey weights and survey design features might cause biased treatment effect estimates. Ignoring survey weights leads to external validity bias, which occurs when people inappropriately make inferences from an unrepresentative analytic sample to a target population. However, there are somewhat conflicting results in this area when using propensity scores with complex survey data, and quite a bit of debate regarding the proper use of complex survey data in general (see e.g., Gelman, 2007, and the associated discussion and rejoinder).

In the propensity score context, based on simulation, Austin et al. (2016) claim that incorporating survey weights into estimates of the propensity scores or not does not make much difference on covariate balance when propensity score matching is used to estimate treatment effects, and DuGoff et al. (2014) argue that because survey design elements such as strata and clustering only affect the variance, and because the propensity score estimation stage only takes the predicted probabilities from the model (not any variance estimates), the propensity score estimation stage does not need to account for the clustering and stratification. However, it is not clear how broadly applicable these results are, the theoretical results in these articles have not been empirically validated, and other authors make other claims.

Recently, Ridgeway et al. (2015) argued that survey weights should be incorporated when estimating propensity scores when propensity score weighting is used to estimate treatment effects.

There is a similar debate in a related research literature examining methods to adjust for survey nonresponse (which uses methods similar to propensity scores in nonexperimental studies), Little and Vartivarian (2003) discussed both weighted and unweighted models for estimating models of survey response. Little and Vartivarian argued that the weighted model is either incorrect or unnecessary. Grau et al. (2006) and Potter et al. (2006) empirically examined Little and Vartivarian's assertion by evaluating the weighted and unweighted logistic regression models to estimate propensity scores for survey nonresponse adjustment using data from the Community Tracking Study Physician Survey. They did not find significant differences in the estimates of variables of interest between the two options.

The purpose of this study is 2-fold: (1) to systematically summarize the appropriate use of survey weights in propensity score analyses of complex survey data and (2) to use a case study to demonstrate the applications of various propensity score analyses of survey data and to empirically compare the population average treatment effect (PATE) estimates using multiple analysis methods. We first review and summarize propensity score methods that handle survey weights in the "Literature Review" section. In the "Case Study: The Effect of Preschool" section, we demonstrate the performance of multiple analysis methods including ordinary least squares (OLS) regression, weighted regression, and various applications of propensity score methods using a nationally representative data set—the ECLS-Kindergarten (ECLS-K)—to estimate the effects of pre-K education on child academic achievement. The correspondence of the results was evaluated using multiple criteria: direction of findings, statistical significance patterns, effect size similarity, statistical difference, and statistical equivalence. "Conclusions" section concludes with a discussion of applications and future research directions.

Literature Review

Propensity score methods average treatment effect (ATE) generally estimate one of two estimands: the ATE or the "average effect of the treatment on the treated" (ATT; Imai et al., 2008; Imbens, 2004; Kurth et al., 2006; McCaffrey et al., 2004; Ridgeway et al., 2012). The ATE is defined as $E[Y(1) - Y(0)]$ and the ATT is defined as $E[Y(1) - Y(0) | Z = 1]$, where $Y(1)$ and $Y(0)$ denote the potential outcomes observed under the treatment ($Z = 1$) and the control group ($Z = 0$). The ATE represents the effect of the

treatment of interest for those in the combined treatment and control groups; the ATT represents the effect for those in just the treatment group (i.e., the average difference in outcomes among treatment group members, comparing their observed outcomes from what would have been expected had they not received the treatment). In a randomized trial, these two estimands are equal in expectation; in a nonexperimental study, they may differ. In situations where the sample at hand is representative of some target population (such as the large-scale surveys referenced above), another important distinction involves sample versus population estimands: The ATT and ATE can each be estimated for the sample or the target population. Imbens (2004) denoted these as SATT/SATE for the sample and PATT/PATE for the population. The choice of estimand affects the approach used and will result in different estimated quantities. Recent work has provided advances in using propensity score methods to analyze survey data to estimate the PATT (e.g., Austin et al., 2016; Lenis et al., 2019; Ridgeway et al., 2015). This article assumes that interest is in estimating the PATE and that propensity score methods will be used to adjust for observed confounders. The usual propensity score methods, such as propensity score subclassification (also called propensity score stratification) and inverse propensity score weighting, can be adapted to estimate PATE and thus will be considered in our study. In addition, the less common approach of regression adjustment for the propensity score (or a function of the propensity score) is used in practice; this approach is included in our study for comparison. Austin (2011) and Stuart (2010) provide reviews of these methods.

Three assumptions are needed for estimating the PATE from the sample (Stuart et al., 2011): (1) All subjects in the population have a positive probability of being selected to receive the treatment given the covariates. (2) Conditional on the observed covariates, the sample selection is independent of the potential outcomes (or, there are no unmeasured variables that are related to both sample selection and the treatment effect). (3) Treatment assignment is independent of sample selection and the potential outcomes, given the observed covariates. See Stuart et al. (2011) for a detailed discussion on these assumptions.

Our review of the empirical education literature that uses propensity score methods to estimate PATE using complex surveys found at least four approaches to handling survey weights at the two stages of propensity score methods—propensity score estimation (Stage 1) and propensity score-based (e.g., subclassification- or weighting-based) causal effect estimation (Stage 2). These approaches include (1) ignoring survey weights in both stages (Hallberg et al., 2011; Hong & Raudenbush, 2006), (2) using survey

weights as a predictor of propensity scores in Stage 1 (Korenman et al., 2012), (3) ignoring survey weights in Stage 1 and Stage 2 subclassifying on the propensity score and estimating treatment effects within subclasses using survey-weighted regression of the outcome (Hahs-Vaughn & Onwuegbuzie, 2006; Zanutto, 2006; Zanutto et al., 2005), and (4) multiplying survey weights and propensity score weights, and using the combined weights in the outcome analysis in Stage 2 but providing no discussion of whether propensity score estimation in Stage 1 used survey weights (Hornik et al, 2001; Schonlau et al, 2004; Zanutto et al., 2005). There is thus a broad array of approaches used in the literature, which is perhaps not surprising given the mixed results in the methodological literature, as noted above, and that before 2014 there was almost no methodological work on the best ways to use propensity scores with complex surveys, with the exception of Zanutto (2006) that compares propensity score and linear regression analysis of complex survey data.

To describe that methodological literature more fully, DuGoff et al. (2014) considered three possible ways of combining propensity scores and complex surveys and used a simple simulation to explore how well those methods estimate population treatment effects. That simulation showed that either not balancing covariates (via propensity score methods) or ignoring survey weights can lead to biased estimates of population treatment effects but that some approaches that combined both propensity score-based covariate balancing and survey weights generated accurate effect estimates. Ridgeway et al. (2015) and Austin et al. (2016) also examined the role of survey weights when estimating propensity scores (Stage 1). Based on mathematical derivation and Monte Carlo simulation, Ridgeway et al. (2015) concluded that propensity scores should be estimated by survey-weighted modeling in stage 1, if Stage 2 uses propensity score weighting to estimating the causal effect. Austin et al. (2016) argued and showed through simulation that using or not using the survey weights in Stage 1 did not matter if Stage 2 uses propensity score matching to estimate the causal effect.

Broadly, there are three options for handling survey weights when estimating propensity scores (Stage 1): (1) “NoWt”: ignoring survey weights, (2) “WtCov”: using the survey weights as a covariate, and (3) “WtModel”: estimating a weighted logistic regression model using the survey weights. Similarly, in the outcome analysis (Stage 2), there are three options: (1) “NoWt”: ignoring survey weights, (2) “WtModel”: estimating a weighted regression model using the original survey weights, and (3) “RWtModel”: estimating a weighted regression model using the reweighted survey weights (i.e., using the product of the propensity score and survey weights). Table 1 summarizes the

Table 1. Summary of Options of Handling Survey Weights in Propensity Score Analyses to Estimate PATE.

| Propensity Score Methods | Label | Stage 1: Estimation (Estimating Propensity Scores) | Stage 2: Use (Outcome Analysis) |
|--------------------------------------|------------------------|--|---------------------------------|
| Using propensity scores as covariate | Cov: NoWt-NoWt | Ignoring | Ignoring |
| | Cov: WtCov-NoWt | Covariate | Ignoring |
| | Cov: NoWt-WtModel | Ignoring | Weighted |
| | Cov: WtModel-WtModel | Weighted | Weighted |
| Subclassification | Sub: NoWt-NoWt | Ignoring | Ignoring |
| | Sub: WtCov-NoWt | Covariate | Ignoring |
| | Sub: NoWt-WtModel | Ignoring | Weighted |
| | Sub: WtModel-WtModel | Weighted | Weighted |
| Propensity score weighting | PSWt: NoWt-NoWt | Ignoring | Ignoring ^a |
| | PSWt: WtCov-NoWt | Covariate | Ignoring ^a |
| | PSWt: NoWt-RWtModel | Ignoring | Reweight |
| | PSWt: WtModel-RWtModel | Weighted | Reweight |

Note. “Reweight” refers to using the product of the original survey weights and propensity score weights as new weights in weighted analysis (Schonlau et al., 2004). PATE = population average treatment effects.

^aIgnoring survey weights but still using the propensity score weights in weighted regression models to estimate PATE.

options for handling survey weights in propensity score analyses that were identified in literature or with natural extension. Note that the total number of possible combinations is 27:3 propensity score estimation methods × 3 propensity score methods (i.e., methods for using propensity scores to estimate the causal effect) × 3 ways of using the survey weights in the propensity score–adjusted samples. Table 1 does not include a complete list of all possible combinations; the ones selected and presented are those that have been used in practice or have the most potential to be used. We use the labels “Method: XX-YY” to identify each combination; “Method” refers to the type of propensity score method, and “XX” and “YY” refer to how survey weights are used in the propensity score estimation stage and in the propensity score use stage (i.e., use in the treatment effect estimate), respectively. We describe these ways of integrating survey weights with propensity score methods in more detail below.

Using Propensity Scores as a Covariate (“Cov”)

This approach for using propensity scores first estimates propensity scores and then fits a regression model of the outcome, with the propensity score

(or a function of the propensity score, e.g., splines) included as a covariate. In this study, we use the logit of the propensity score. There are four plausible ways to incorporate survey weights into this propensity score method to estimate the PATE: (1) “Cov: NoWt-NoWt,” that is, ignoring the survey weights entirely from both the estimation and use stages of propensity score methods (e.g., Hallberg et al., 2011; Hong & Raudenbush, 2006), (2) “Cov: WtCov-NoWt,” that is, using the survey weight as a covariate at the estimation stage but ignoring the survey weights at the use stage (e.g., Korenman et al., 2012), (3) “Cov: NoWt-WtModel,” that is, ignoring the survey weights at the estimation stage but fitting a weighted regression model at the use stage, and (4) “Cov: WtModel-WtModel,” that is, using survey weights to do weighted analysis at both estimation and use stages.

Subclassification (“Sub”)

Propensity score subclassification refers to a method that groups individuals into subclasses (e.g., deciles) defined by the propensity score (we use the logit of the propensity score in this study). Effect estimates are obtained by estimating an effect within each subclass and then averaging across subclasses. In our survey context, “Sub: NoWt-NoWt” and “Sub: WtCov-NoWt” both group individuals into propensity score subclasses; both do not utilize the survey weights in the propensity score use stage; the former also doesn’t use the survey weights in the propensity score estimation, whereas the latter does use the survey weights as a predictor in the propensity score model. “Sub: NoWt-WtModel” refers to subclassification on the propensity score estimated by an unweighted logistic regression model and then estimating treatment effects within each subclass using a survey-weighted regression of the outcome (e.g., Hahs-Vaughn & Onwuegbuzie, 2006; Zanutto, 2006; Zanutto et al., 2005). PATE is estimated by weighting the subclass-specific treatment effects by the proportion (p_j) of the population in that subclass: $p_j = \frac{N_{tj} + N_{cj}}{\sum_j (N_{tj} + N_{cj})}$, where N_{tj} and N_{cj} are the

numbers of subjects in the treatment and control groups for subclass j in the total population (which can be estimated using a sum of the survey weights). Note that in “Sub: NoWt-NoWt” and “Sub: WtCov-NoWt,” the proportions in each subclass are calculated using the sample sizes not the population sizes. “Sub: WtModel-WtModel” is similar to “Sub: NoWt-WtModel” but fits a weighted propensity score model.

Propensity Score Weighting (“PSWt”)

This category of propensity score methods uses the inverse probability of treatment weighting to estimate the PATE. Specifically, the propensity score weights for PATE are $w_i = \frac{T_i}{\hat{e}_i} + \frac{1-T_i}{1-\hat{e}_i}$, where, T_i is a binary variable indicating the treatment status ($T = 1$ for the treatment group and $T = 0$ for the comparison group), and \hat{e}_i is the estimated propensity score for individual i . “PSWt: NoWt-NoWt” and “PSWt: WtCov-NoWt” are analogous to similar methods described above. “PSWt: NoWt-RWtModel” and “PSWt: WtModel-RWtModel” methods utilize a combined weight in the outcome model, which is the product of the survey weights and propensity score weights (w_i ; Hornik et al, 2001; Ridgeway et al., 2015; Schonlau et al., 2004). Note that “RWtModel” refers to the use of the combined weights, while “WtModel” (also associated with propensity score subclassification and propensity score as covariate methods) refers to the use of the survey weights.

In all outcome analyses (Stage 2), covariates can be included in analytic models to further reduce selection bias and improve statistical power.

Case Study: The Effect of Preschool

The purpose of this case study is to empirically compare the PATE estimates using multiple analysis methods that include OLS regression, weighted least square (WLS) regression, and various propensity score applications. The empirical question for the applied researchers is “how much difference it would make if different propensity score applications are selected to analyze the survey data to estimate the population average treatment effect?” This case study is to estimate the PATE of preschool as compared to parental care for children across the U.S. on child math achievement.

Background

To close the achievement gap, researchers such as Ladd (2012) proposed several policy interventions, including early childhood care and education (ECCE) programs, such as center-based programs like preschool, prekindergarten, and Head Start. Moreover, the effects of home-based care settings such as care by a relative or nonrelative (i.e., kith & kin care or family childcare) as well as parental care have been of interest. The reason ECCE programs are lauded as a method for preventing the achievement gap at

school entry and beyond is that the effectiveness of these programs has been widely demonstrated. Most ECCE studies investigated whether ECCE programs, either center-based (Magnuson et al., 2007) or home-based (Fuller et al., 2004), improved children's academic achievement. Although some studies indicate the positive effects of ECCE, several studies indicate that the positive effects of ECCE were mixed (Barnett, 2011; Lipsey et al., 2013). The policy question to be answered in this case study is "What would be the average difference in math achievement scores at Kindergarten entry if all children in the US went to a formal preschool before kindergarten, as compared to all receiving parental care?"

Data

The Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K) is a nationally representative longitudinal study of children. The ECLS-K used a multistage, stratified, clustered sampling design (U.S. Department of Education, National Center for Education Statistics, 2009). Counties or groups of contiguous counties were separated into primary sampling units (PSUs). There were 100 PSUs selected for the ECLS-K in Stage 1. The PSUs were first stratified based on selected characteristics such as region, median household income, and metropolitan status, and then the PSUs were sampled within strata with probability proportional to size. At Stage 2, schools were sampled within sampled PSUs and there were 1,277 schools sampled. At Stage 3, children were sampled with schools and there were a total of 22,666 children attending kindergarten during the 1998–99 school year sampled. The ECLS-K includes an oversample of Asian/Pacific Islander children. Survey weights are used to make the sample data representative of the target population and account for differential selection probabilities and differential patterns of response/nonresponse. Direct cognitive assessments on math and reading were administered in the fall of kindergarten through the spring of Grade 8. Additional extensive data were collected at multiple time points, including kindergarten entry, including child and family characteristics.

Following Magnuson et al. (2007), to define the treatment and comparison groups we used the parental response to the fall kindergarten survey question "primary type non-parental care at prekindergarten" (variable P1PRIMPK; U.S. Department of Education, National Center for Education Statistics, 2009). This led to two groups of interest: center-based preschool ("treatment"; $N = 7,367$) and parental care ("comparison"; $N = 3,150$).

The outcome variable considered is children's math achievement in the fall of kindergarten, which is the item response theory (IRT) scale score constructed by the U.S. Department of Education's National Center for Education Statistics (NCES). The outcome measure has high reliability, with a Cronbach's α coefficient of .88 (Tourangeau et al., 2009). In the logistic regression model used to estimate the propensity scores, we included the covariates that were correlated either with the outcome or the treatment status. Following Magnuson et al. (2007), covariates included in the propensity score models were gender, race, height, weight, age at the kindergarten entry, English-speaking status at home, parents' educational level, income, composite socioeconomic status (SES) measure, household structure (numbers of parents and siblings), census region in sampling frame (northeast, midwest, south, and west), and locality (urban, rural, or town). Table 2 presents the descriptive statistics of the outcome variable and covariates by condition. The results of both unweighted analysis and survey-weighted analysis are reported. Note that there are more Hispanic children, less children speaking English at home, children with lower motor skills, lower family income, lower parent highest education level, and lower SES in parental care than in preschool. The survey weight used is the child weight for the full sample at the fall of kindergarten (variable C1CW0), which allows us to adjust for nonresponse to every component of every round of data collection.

Method

Data analysis methods. The data analysis methods that we compare for estimating the PATE of preschool versus parental care include the OLS regression ignoring survey weights, design-based survey-WLS regression, and the 12 propensity score methods that were listed in Table 1. We include OLS regression, a naive approach to analyzing survey data, just as a comparison. The WLS regression that uses the survey weights and design features such as strata is commonly used for analyzing survey data. Hence, we use the PATE estimate from the WLS regression as the reference group and compare the PATE estimates of all the other analysis methods with the WLS regression. At Stage 1 of the 12 propensity score analyses, we use the (weighted or unweighted) logistic regression models to estimate the propensity score using 22 covariates listed in Table 2. In all our outcome analyses at Stage 2, we included the 22 covariates in our analytic models to further reduce selection bias and improve statistical power. All the standard error estimates are model based.

Table 2. Descriptive Statistics of the Outcome Variable and Covariates by Conditions.

| Variables | Unweighted Analysis | | | | Weighted Analysis | | | |
|--------------------------------|---------------------|--------|--------|---------------------------------|-------------------|------|-----------|---------------------------------|
| | Parental Care | | Pre-K | Standardized Mean Difference | Parental Care | | Pre-K | Standardized Mean Difference |
| | Mean | SD | Mean | SD | Mean | SE | Mean | SE |
| Math score in Fall K (outcome) | 23.92 | 8.29 | 28.71 | 9.73 | 23.66 | 0.16 | 28.37 | 0.12 |
| Female | 0.48 | 0.50 | 0.49 | 0.50 | 0.47 | 0.01 | 0.49 | 0.01 |
| Black | 0.10 | 0.30 | 0.11 | 0.32 | 0.11 | 0.01 | 0.12 | 0.00 |
| Hispanic | 0.27 | 0.44 | 0.12 | 0.33 | 0.28 | 0.01 | 0.13 | 0.00 |
| Other race | 0.11 | 0.32 | 0.09 | 0.28 | 0.07 | 0.00 | 0.06 | 0.00 |
| Northeast | 0.17 | 0.38 | 0.20 | 0.40 | 0.16 | 0.01 | 0.20 | 0.00 |
| Midwest | 0.20 | 0.40 | 0.25 | 0.43 | 0.19 | 0.01 | 0.24 | 0.01 |
| South | 0.33 | 0.47 | 0.34 | 0.48 | 0.37 | 0.01 | 0.37 | 0.01 |
| Town | 0.36 | 0.48 | 0.44 | 0.50 | 0.40 | 0.01 | 0.49 | 0.01 |
| Rural | 0.21 | 0.41 | 0.15 | 0.36 | 0.21 | 0.01 | 0.15 | 0.00 |
| Two parents | 0.81 | 0.39 | 0.82 | 0.38 | 0.81 | 0.01 | 0.81 | 0.00 |
| Two parents with siblings | 0.75 | 0.43 | 0.70 | 0.46 | 0.75 | 0.01 | 0.70 | 0.01 |
| Two parents without siblings | 0.07 | 0.25 | 0.12 | 0.32 | 0.06 | 0.00 | 0.11 | 0.00 |
| One parents with siblings | 0.13 | 0.34 | 0.11 | 0.31 | 0.13 | 0.01 | 0.12 | 0.00 |
| One parents without siblings | 0.04 | 0.19 | 0.06 | 0.24 | 0.03 | 0.00 | 0.06 | 0.00 |
| Biological mother | 0.95 | 0.21 | 0.95 | 0.22 | 0.95 | 0.00 | 0.95 | 0.00 |
| Speaking English at home | 0.79 | 0.41 | 0.92 | 0.28 | 0.79 | 0.01 | 0.92 | 0.00 |
| Height (inches) | 44.53 | 2.18 | 44.77 | 2.15 | 44.52 | 0.04 | 44.76 | 0.03 |
| Weight (pounds) | 45.95 | 8.87 | 46.33 | 8.27 | 45.95 | 0.17 | 46.35 | 0.10 |
| Age (months) | 65.40 | 4.44 | 65.69 | 4.22 | 65.48 | 0.08 | 65.72 | 0.05 |
| Family income (\$) | 40,365 | 40,127 | 66,879 | 64,079 | 39,235 | 710 | 65,144 | 739 |
| Parent highest education level | 4.12 | 1.90 | 5.40 | 1.88 | 4.01 | 0.04 | 5.31 | 0.02 |
| SES | -0.24 | 0.75 | 0.31 | 0.76 | -0.29 | 0.01 | 0.27 | 0.01 |
| N | 3,150 | | 7,367 | | 3,150 | | 7,367 | |
| Sum of weights | | | | | 650,612 | | 1,457,650 | |

Note. The absolute values of standardized mean differences bigger than .25 are in boldface. SES = socioeconomic status; SD = standard deviation; SE = standard error.

To assess the similarity between exposed and unexposed groups, we checked the covariate balance. Covariates are more balanced in all the propensity score analyses than in the OLS or WLS analyses. For example, Table A1 in the appendix presents the covariate balance checking results for four applications of propensity score weighting (“PSWt”). None of 22 covariates has a standardized mean difference bigger than .50, only one covariate (family income) has a standardized mean difference between .25 and .50, and all other covariates have standardized mean differences smaller than .15 in all four applications of PSWt, while two covariates (SES and parental highest education level) have standardized mean differences bigger than .50 and three covariates (family income, Hispanic, speaking English at home, and motor skills) have standardized mean differences between .25 and .50 in the OLS and WLS analyses before using propensity score methods (Table 2). Figure 1 presents the covariate balance checking results for the WLS analysis before using the propensity score methods and after using “PSWt: WtModel-RWtModel.” The covariate balance checking results suggest that propensity score methods did reduce selection bias due to observed variables, relative to the original sample (which is used for OLS and WLS methods). In addition, there is not much difference on the covariate balance checking results among four applications of PSWt.

Assessing the correspondence of the PATE estimates between different methods. Traditionally, researchers compared different methods using null hypothesis statistical testing. Although null hypothesis statistical testing is useful in testing if there is any statistically significant difference between estimates, it is not able to draw conclusions on the similarity or equivalence of the estimates. Recently, multiple criteria were used for assessing the correspondence of the results between different methods in the within-study comparison literature (Cook et al., 2008; Steiner & Wong, 2018; Wilde & Hollister, 2007). The range of these criteria is from very liberal to very strict, which include whether the different estimates are in the same direction (positive or negative), whether the analyses reached the same conclusion about statistical significance, the absolute difference between the two effect sizes obtained, the relative difference between the two effect sizes (the percentage of the absolute difference divided by the effect size of the reference estimate, and multiplied by 100), the null hypothesis statistical significance test, and the equivalence test.

Below we review the *inferential confidence interval* (ICI) for the statistical significance test and equivalence test. Conventional confidence intervals represent the uncertainty associated with each effect estimate individually and can

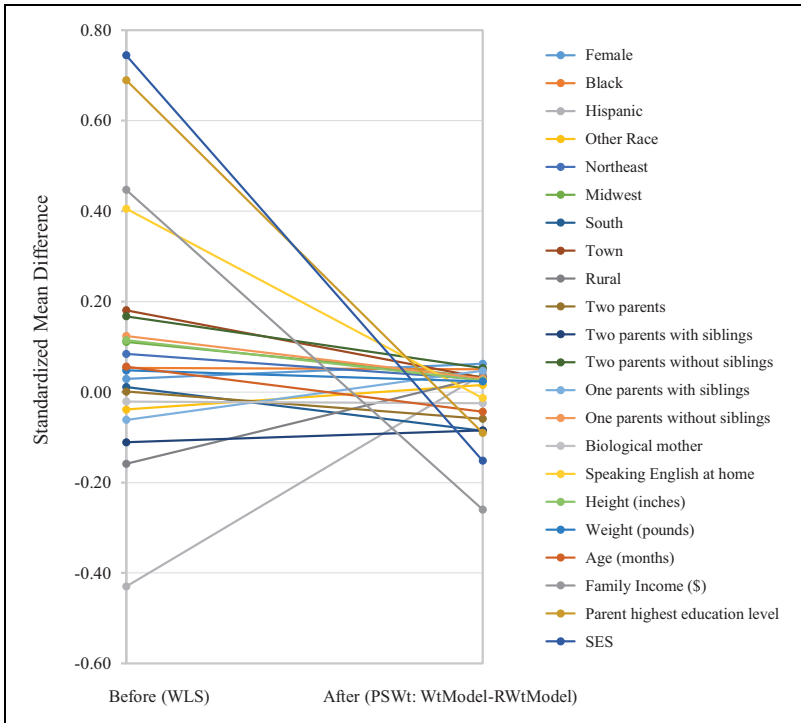


Figure 1. Covariate balance checking before and after propensity score weighting. Note. Weighted least square was used before applying propensity score methods; “PSWt: WtModel-RWtModel” was used for the propensity score weighting.

overlap even when there is a statistically significant difference between the estimates (Cumming & Finch, 2005). The ICI method proposed by Tryon (2001; Tryon & Lewis, 2008) addresses this problem by creating confidence intervals around the estimates from which statistical significance can be assessed directly on the basis of their overlap. The ICIs for Parameter 1 (θ_1) and Parameter 2 (θ_2) are $(\hat{\theta}_1 - Et_{v_1}^{\alpha/2} S_1, \hat{\theta}_1 + Et_{v_1}^{\alpha/2} S_1)$ and $(\hat{\theta}_2 - Et_{v_2}^{\alpha/2} S_2, \hat{\theta}_2 + Et_{v_2}^{\alpha/2} S_2)$, respectively, where $\hat{\theta}_1$ and $\hat{\theta}_2$ are the point estimates of Parameters 1 and 2, S_1 and S_2 are their estimated standard errors, $t_{v_1}^{\alpha/2}$ and $t_{v_2}^{\alpha/2}$ are t values with v_1 and v_2 degrees of freedom at a specified α level (α), and E is a reduction factor that makes the t test of the difference between the parameters just significant when the upper bound of one ICI and the lower bound of the other are equal. This reduction factor is defined as $E = \frac{t_{v_2}^{\alpha/2} S_2 - 1}{t_{v_1}^{\alpha/2} S_1 + t_{v_2}^{\alpha/2} S_2}$ (Tryon &

Lewis, 2008, equation 9), where $t_{v12}^{\alpha/2}$ is the t value for the difference between two parameters with $v12 = v1 + v2$ degrees of freedom; and S_{2-1} is the estimated standard error of the difference between the two parameters. In addition, because we have multiple estimates to assess, there is a multiple comparisons issue that increases the potential for chance findings of statistical significance. To address this, we followed Goldstein and Healy's (1995) recommendation to create ICI limits quantitatively and graphically to compare multiple treatment effect estimates based on average E values across all planned comparisons (see Tryon, 2001, for more discussion of this issue).

This use of the ICI allows us to address the question of whether there is a statistically significant difference between multiple PATE estimates. As such, however, it is based on null hypothesis testing with its inherent orientation toward testing whether two estimates are different (i.e., rejecting the null hypothesis). Failing to reject the null hypothesis is not the same as affirming that the two estimates are statistically equivalent. Wong and Steiner (2018) suggested that equivalence tests be used for this latter purpose. The hypothesis for a test of statistical equivalence is that the difference between two parameter estimates is Δ , where Δ is "an amount that is considered to be inconsequential on the substantive grounds that have been established apart from the analysis at hand by professional consensus or other means" (Tryon & Lewis, 2008, p. 274). The ICI analysis described above can also be used for equivalence testing (Tryon & Lewis, 2008). In general form, it is algebraically equivalent to another method of testing for statistical equivalence using two one-sided t tests (Schuirmann, 1987). Within the ICI framework, statistical equivalence is affirmed if the ICIs overlap and $eRg \leq \Delta$, where eRg is the *equivalence range*, the interval between the lower ICI limit of the smaller parameter estimate and the upper ICI limit of the larger parameter estimate. The equivalence range is thus the most of the two parameter estimates can differ while still falling within their respective ICIs. If that range is no greater than Δ , the estimates are judged to be statistically equivalent.

There is no consensus in the within-study comparisons or replication studies literature about how close the effect estimates among different methods should be considered equivalent. A recent study by Steiner and Wang (2018) suggested using .10 standard deviation as the reasonable tolerance threshold while Dong and Lipsey (2018) suggested a threshold effect size of .05 if the reference effect size is .25 or less and a threshold of 20% of the reference effect size if it is bigger than .25. In this study, we conducted an equivalence test using the tolerance thresholds suggested by both Dong and Lipsey (2018) and Steiner and Wang (2018).

Results

Table 3 presents a summary of the estimates of the PATE (Cohen's d) of preschool on children's kindergarten math scores. The standard error of the PATE estimates was model based. First, all the analyses produced significant PATE estimates (ranging from .17 to .22 standard deviations) favoring preschool as compared to family care. Second, compared to the WLS estimate, the PATE estimates from the propensity score-based methods have absolute differences ranging from 0.01 to 0.05 standard deviations and relative differences ranging from 7.9% to 25.9%. Third, including the survey weights in the outcome analysis (WLS, "Methods: NoWt-WtModel," and "Methods: WtModel-WtModel") produced slightly smaller PATE estimates, and the propensity score methods produced slightly bigger PATE estimates than the OLS and WLS analyses. However, by checking the overlap of the 95% ICIs of the PATE estimates (Figure 2), there is no statistical difference among these methods. Fourth, the equivalence range (eRg) ranges from 0.07 to 0.11. Using the tolerance threshold ($\Delta = .10$) suggested by Steiner and Wang (2018), we could conclude that "PSWt: NoWt-NoWt" and "PSWt: WtCov-NoWt" are not statistically equivalent to the WLS analysis and must be characterized as statistically indeterminate while all the other analyses are statistically equivalent to the WLS analysis. For example, "PSWt: NoWt-RWtModel" and "PSWt: WtModel-RWtModel" that took account of survey weights in the weighted analysis of the PATE are statistically equivalent to the WLS analysis. However, using the tolerance threshold ($\Delta = .05$) suggested by Dong and Lipsey (2018), we could conclude that all the analyses are statistically indeterminate because all eRg are bigger than the tolerance threshold ($\Delta = .05$), and there is no significant difference among these analyses.

Conclusions

In the large body of literature about propensity score methods, very few studies concern survey weights with the exception of Austin et al. (2016), DuGoff et al. (2014), Lenis et al. (2019), Hornik et al (2001), Ridgeway et al. (2015), and Zanutto (2006). However, these studies only focused on certain propensity score applications and did not compare a variety of propensity score methods for estimating the PATE. This study attempts to summarize and provide suggestions about how to handle survey weights in propensity score analyses of survey data from complicated sampling design, when interest is in estimating the PATE. In particular, we conducted

Table 3. Correspondence Assessed With Multiple Criteria on the PATE Estimates Between the Weighted Regression Analysis and the Other Analyses.

| Analytic Methods | Effect Size | p Value | Lower 95% ICI ^a | Upper 95% ICI ^b | eRg ^c | Statistical Difference ^d | Statistical Equivalence ($\Delta = .05$) | Statistical Equivalence ($\Delta = .10$) | Absolute Difference ^e | Relative Difference ^f (%) |
|------------------------|-------------|-----------------|----------------------------|----------------------------|------------------|-------------------------------------|--|--|----------------------------------|--------------------------------------|
| WLS | .17 | <.001 | .15 | .20 | NA | NA | NA | NA | NA | NA |
| OLS | .19 | <.001 | .16 | .21 | .07 | no | no | yes | .01 | 8.0 |
| Cov: NoWt-NoWt | .20 | <.001 | .17 | .23 | .09 | no | no | yes | .03 | 16.4 |
| Cov: WtCov-NoWt | .20 | <.001 | .17 | .23 | .08 | no | no | yes | .03 | 15.4 |
| Cov: NoWt-WtModel | .19 | <.001 | .16 | .22 | .07 | no | no | yes | .01 | 7.9 |
| Cov: WtModel-WtModel | .19 | <.001 | .16 | .22 | .07 | no | no | yes | .01 | 8.0 |
| Sub: NoWt-NoWt | .20 | <.001 | .17 | .24 | .09 | no | no | yes | .03 | 17.5 |
| Sub: WtCov-NoWt | .20 | <.001 | .17 | .24 | .09 | no | no | yes | .03 | 17.5 |
| Sub: NoWt-WtModel | .19 | <.001 | .15 | .22 | .08 | no | no | yes | .01 | 8.4 |
| Sub: WtModel-WtModel | .19 | <.001 | .16 | .23 | .08 | no | no | yes | .02 | 9.0 |
| PSWt: NoWt-NoWt | .22 | <.001 | .19 | .25 | .11 | no | no | no | .05 | 25.9 |
| PSWt: WtCov-NoWt | .22 | <.001 | .19 | .25 | .11 | no | no | no | .04 | 25.4 |
| PSWt: NoWt-RWtModel | .20 | <.001 | .17 | .24 | .09 | no | no | yes | .03 | 16.9 |
| PSWt: WtModel-RWtModel | .20 | <.001 | .17 | .24 | .09 | no | no | yes | .03 | 17.0 |

Note. PATE = population average treatment effects; WLS = weighted least square; OLS = ordinary least square; NA = not applicable.

^aLower 95% ICI is the lower limit of the 95% inferential confidence interval. ^bUpper 95% ICI is the upper limit of the 95% inferential confidence interval. ^ceRg is the equivalence range (Tryon & Lewis, 2008). ^dStatistical difference indicates whether the difference between the weighted regression analysis and the other analysis is statistically significant. ^eAbsolute difference is calculated as the effect size difference between the other analysis and the weighted regression analysis (.17). ^fPercentage of bias is calculated as $100 \times (\text{Absolute difference}/.17)$.

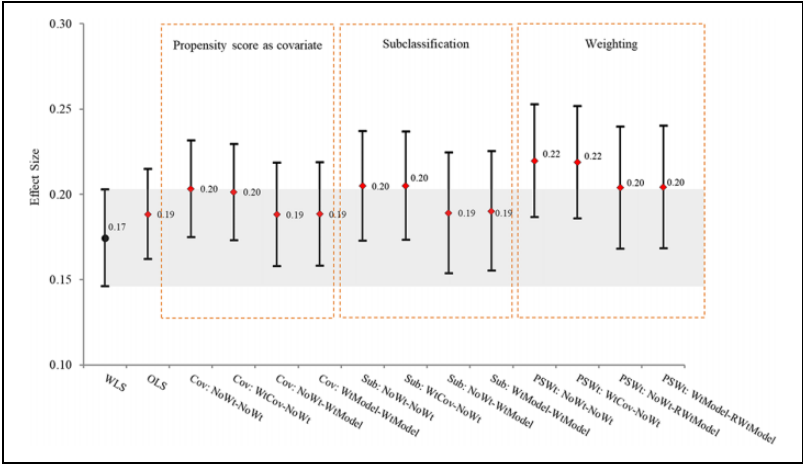


Figure 2. Population average treatment effects estimates and 95% inferential confidence intervals for the weighted least square analysis and the other analyses. *Note.* The gray area covers the 95% inferential confidence intervals (.09 to .15) of the effect size estimate using the weighted regression analysis.

a case study to empirically compare 12 propensity score methods for estimating the population average effect of preschool on children’s math achievement in kindergarten. Some conclusions can be drawn as follows:

1. The literature suggests that using survey weights or not when estimating propensity scores (Stage 1) leads to similar results in the PATE for two propensity score methods: propensity score subclassification (Austin et al., 2016) and propensity score matching (Lenis et al., 2019). However, Ridgeway et al. (2015) suggested survey-weighted analysis to estimate the propensity scores to protect against model misspecification when the propensity score weighting method is used. We recommend using survey weights at the propensity score estimation stage in all propensity score analyses.
2. When using the propensity score to estimate the PATE (Stage 2), survey weights must be taken into account in the outcome model to yield accurate PATE estimates for all propensity score methods (Austin et al., 2016; DuGoff et al., 2014; Lenis et al., 2019; Ridgeway et al., 2015). This is easy to understand since the survey weights are critical for making inferences to the target population.

3. In our case study to estimate the population effects of preschool, three categories of propensity score methods (using propensity score as covariate, propensity score subclassification, and propensity score weighting) have similar performance in reducing covariance imbalance and produced similar PATE estimates. There is no statistical difference among the PATE estimates based on WLS and the 12 propensity score methods. All the analyses conclude that preschool had a significant positive effect on improving children's math achievement in kindergarten. However, they are statistically equivalent only when the tolerance threshold (Δ) is big, for example, $\Delta = .10$, and they are statistically indeterminate when Δ is small, for example, $\Delta = .05$.

In summary, it is important for researchers to think carefully about their estimand of interest and use methods appropriate for that estimand. If interest is in drawing inferences to the survey target population (i.e., in estimating the PATE), it is important to take the survey weights into account. Propensity score methods are useful for reducing bias due to imbalanced covariates in estimating PATE. Although our case study produced mixed results regarding equivalence test depending on the magnitude of the tolerance threshold (Δ), it would be safe to include survey weights in the weighted analysis in both estimation and use stages of three applications of propensity score methods (use propensity score as a covariate, propensity score subclassification, and propensity score weighting) for estimating the PATE. We hope that this article raises awareness of this issue and provides concrete tools and recommendations for researchers to use when estimating causal effects in populations represented by sample surveys.

Appendix

Table A1. Covariates Balance Checking.

| Variables | PSWt: NoWt-NoWt | | | | | | PSWt: WtGov-NoWt | | | | | | PSWt: NoWt-RWtModel | | | | | | PSWt: WtModel-RWtModel | | | | | |
|--------------------------------|------------------------------|---------|--------|--------|--------|--------|------------------|---------|--------|--------|--------|--------|---------------------|-----------|-----------|-----------|-----------|-----------|------------------------|-----------|-----------|-----------|-----------|-----------|
| | Parental Care | | | Pre-K | | | Parental Care | | | Pre-K | | | Parental Care | | | Pre-K | | | Parental Care | | | Pre-K | | |
| | Standardized Mean Difference | | | Mean | | | Mean | | | SD | | | Mean | | | SD | | | Mean | | | SD | | |
| | Mean | SD | SE | Mean | SD | SE | Mean | SD | SE | Mean | SD | SE | Mean | SD | SE | Mean | SD | SE | Mean | SD | SE | Mean | SD | SE |
| Female | 0.43 | 0.98 | 0.49 | 0.49 | 0.60 | 0.07 | 0.43 | 0.98 | 0.49 | 0.49 | 0.60 | 0.07 | 0.43 | 0.94 | 0.04 | 0.43 | 0.94 | 0.04 | 0.43 | 0.94 | 0.04 | 0.48 | 0.01 | .06 |
| Black | 0.09 | 0.56 | 0.11 | 0.37 | 0.05 | 0.05 | 0.09 | 0.56 | 0.11 | 0.37 | 0.05 | 0.05 | 0.09 | 0.01 | 0.01 | 0.12 | 0.00 | 0.01 | 0.09 | 0.01 | 0.12 | 0.00 | 0.00 | .05 |
| Hispanic | 0.14 | 0.69 | 0.17 | 0.45 | 0.04 | 0.04 | 0.14 | 0.69 | 0.17 | 0.45 | 0.04 | 0.04 | 0.15 | 0.02 | 0.01 | 0.18 | 0.01 | 0.15 | 0.02 | 0.18 | 0.01 | 0.18 | 0.01 | .04 |
| Other Race | 0.09 | 0.57 | 0.09 | 0.35 | 0.00 | 0.00 | 0.09 | 0.57 | 0.09 | 0.35 | 0.00 | 0.00 | 0.06 | 0.01 | 0.06 | 0.06 | 0.00 | 0.02 | 0.06 | 0.01 | 0.06 | 0.00 | 0.02 | .02 |
| Northeast | 0.18 | 0.76 | 0.19 | 0.47 | 0.01 | 0.01 | 0.18 | 0.76 | 0.19 | 0.47 | 0.01 | 0.01 | 0.17 | 0.02 | 0.19 | 0.01 | 0.19 | 0.01 | 0.17 | 0.02 | 0.19 | 0.01 | 0.01 | .03 |
| Midwest | 0.21 | 0.81 | 0.24 | 0.51 | 0.03 | 0.03 | 0.22 | 0.81 | 0.24 | 0.51 | 0.03 | 0.03 | 0.20 | 0.02 | 0.22 | 0.01 | 0.22 | 0.01 | 0.20 | 0.02 | 0.22 | 0.01 | 0.01 | .03 |
| South | 0.40 | 0.97 | 0.34 | 0.57 | -0.08 | -0.08 | 0.40 | 0.97 | 0.34 | 0.57 | -0.08 | -0.08 | 0.44 | 0.06 | 0.37 | 0.01 | 0.37 | 0.01 | 0.44 | 0.06 | 0.37 | 0.01 | 0.01 | -.09 |
| Town | 0.39 | 0.97 | 0.41 | 0.59 | 0.03 | 0.03 | 0.39 | 0.96 | 0.41 | 0.59 | 0.03 | 0.03 | 0.43 | 0.04 | 0.46 | 0.01 | 0.46 | 0.01 | 0.43 | 0.04 | 0.46 | 0.01 | 0.01 | .03 |
| Rural | 0.15 | 0.70 | 0.17 | 0.45 | 0.04 | 0.04 | 0.15 | 0.70 | 0.17 | 0.45 | 0.04 | 0.04 | 0.15 | 0.02 | 0.17 | 0.01 | 0.17 | 0.01 | 0.15 | 0.02 | 0.17 | 0.01 | 0.01 | .03 |
| Two parents | 0.85 | 0.70 | 0.82 | 0.46 | -0.06 | -0.06 | 0.85 | 0.70 | 0.82 | 0.46 | -0.06 | -0.06 | 0.85 | 0.02 | 0.81 | 0.01 | 0.81 | 0.01 | 0.85 | 0.02 | 0.81 | 0.01 | 0.01 | -.06 |
| Two parents with siblings | 0.78 | 0.82 | 0.72 | 0.54 | -0.08 | -0.08 | 0.78 | 0.82 | 0.72 | 0.54 | -0.08 | -0.08 | 0.77 | 0.02 | 0.71 | 0.01 | 0.71 | 0.01 | 0.77 | 0.02 | 0.71 | 0.01 | 0.01 | -.08 |
| Two parents without siblings | 0.08 | 0.53 | 0.10 | 0.36 | 0.05 | 0.05 | 0.08 | 0.53 | 0.10 | 0.36 | 0.05 | 0.05 | 0.07 | 0.01 | 0.10 | 0.00 | 0.10 | 0.00 | 0.07 | 0.01 | 0.10 | 0.00 | 0.00 | .05 |
| One parent with siblings | 0.09 | 0.57 | 0.11 | 0.38 | 0.05 | 0.05 | 0.09 | 0.57 | 0.11 | 0.38 | 0.05 | 0.05 | 0.10 | 0.01 | 0.12 | 0.00 | 0.12 | 0.00 | 0.10 | 0.01 | 0.12 | 0.00 | 0.00 | .03 |
| One parent without siblings | 0.04 | 0.40 | 0.05 | 0.27 | 0.03 | 0.03 | 0.04 | 0.40 | 0.05 | 0.27 | 0.03 | 0.03 | 0.04 | 0.01 | 0.05 | 0.00 | 0.05 | 0.00 | 0.04 | 0.01 | 0.05 | 0.00 | 0.00 | .03 |
| Biological mother | 0.96 | 0.38 | 0.95 | 0.26 | -0.03 | -0.03 | 0.96 | 0.38 | 0.95 | 0.26 | -0.03 | -0.03 | 0.96 | 0.01 | 0.95 | 0.00 | 0.95 | 0.00 | 0.96 | 0.01 | 0.95 | 0.00 | 0.00 | -.02 |
| Stepfather | 0.88 | 0.63 | 0.88 | 0.39 | -0.02 | -0.02 | 0.88 | 0.63 | 0.88 | 0.39 | -0.02 | -0.02 | 0.89 | 0.01 | 0.89 | 0.01 | 0.89 | 0.01 | 0.89 | 0.01 | 0.89 | 0.01 | 0.01 | -.01 |
| Speaking English at home | 44.61 | 4.10 | 44.70 | 2.57 | 0.02 | 0.02 | 44.62 | 4.10 | 44.70 | 2.57 | 0.02 | 0.02 | 44.60 | 0.07 | 44.69 | 0.03 | 44.69 | 0.03 | 44.60 | 0.07 | 44.69 | 0.03 | 0.03 | .03 |
| Height (inches) | 45.97 | 16.50 | 46.24 | 9.96 | 0.02 | 0.02 | 45.99 | 16.53 | 46.23 | 9.96 | 0.02 | 0.02 | 45.92 | 0.30 | 46.24 | 0.11 | 46.24 | 0.11 | 45.92 | 0.30 | 46.24 | 0.11 | 0.11 | .02 |
| Weight (pounds) | 65.94 | 8.33 | 65.60 | 5.04 | -0.05 | -0.05 | 65.93 | 8.33 | 65.60 | 5.04 | -0.05 | -0.05 | 65.93 | 0.32 | 65.63 | 0.06 | 65.63 | 0.06 | 65.93 | 0.32 | 65.63 | 0.06 | 0.06 | -.04 |
| Age (months) | 154.635 | 482.157 | 59.107 | 68.767 | -0.27 | -0.27 | 151.977 | 475.563 | 59.102 | 68.775 | -0.26 | -0.26 | 149.166 | 62.699 | 57.454 | 646 | 646 | 646 | 149.166 | 62.699 | 57.454 | 646 | 646 | -.26 |
| Family income (\$) | 5.39 | 4.15 | 5.03 | 2.31 | -0.10 | -0.10 | 5.38 | 4.14 | 5.03 | 2.31 | -0.10 | -0.10 | 5.24 | 0.18 | 4.93 | 0.03 | 4.93 | 0.03 | 5.24 | 0.18 | 4.93 | 0.03 | 0.03 | -.09 |
| Parent highest education level | 0.39 | 1.86 | 0.14 | 0.98 | -0.16 | -0.16 | 0.38 | 1.85 | 0.14 | 0.98 | -0.16 | -0.16 | 0.33 | 0.14 | 0.10 | 0.02 | 0.10 | 0.02 | 0.33 | 0.14 | 0.10 | 0.02 | 0.02 | -.15 |
| N | 3,150 | 7,367 | 10,509 | 10,509 | 10,509 | 10,509 | 3,150 | 7,367 | 10,509 | 10,509 | 10,509 | 10,509 | 3,150 | 7,367 | 10,509 | 3,150 | 7,367 | 10,509 | 3,150 | 7,367 | 10,509 | 3,150 | 7,367 | 7,367 |
| Sum of weights | 11,704 | 11,704 | 11,704 | 11,704 | 11,704 | 11,704 | 12,279 | 12,279 | 12,279 | 12,279 | 12,279 | 12,279 | 2,463,412 | 2,463,412 | 2,093,599 | 2,463,412 | 2,093,599 | 2,093,599 | 2,463,412 | 2,093,599 | 2,093,599 | 2,093,599 | 2,093,599 | 2,093,599 |

Note. The absolute values of standardized mean differences bigger than .25 are in boldface. SES = socioeconomic status; SD = standard deviation; SE = standard error.

Authors' Note

David Lenis is now affiliated with Covera Health, New York, NY, USA.


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: This research was supported by the National Institute of Mental Health (K25MH083846, R01MH099010, Principal Investigator Stuart) and the U.S. Department of Education, Institute of Education Sciences (R305D150001, Principal Investigators Stuart and Dong).

ORCID iD

Nianbo Dong  <https://orcid.org/0000-0003-0128-7106>

References

- Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46*(3), 399–424. <https://doi.org/10.1080/00273171.2011.568786>
- Austin, P. C., Jembere, N., & Chiu, M. (2016). Propensity score matching and complex surveys. *Statistical Methods in Medical Research, 27*, 1240–1257.
- Barnett, W. S. (2011). Effectiveness of early educational intervention. *Science, 333*(6045), 975–978.
- Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies often produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management, 27*(4), 724–750.
- Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals, and how to read pictures of data. *American Psychologist, 60*, 170–180.
- Dong, N., & Lipsey, M. W. (2018). Can propensity score analysis approximate experiments using pretest and demographic information in Pre-K intervention research? *Evaluation Review, 42*(1), 34–70. <https://doi.org/10.1177/0193841X17749824>
- DuGoff, E. H., Schuler, M., & Stuart, E. A. (2014). Generalizing observational study results: Applying propensity score methods to complex surveys. *Health Services Research, 49*(1), 284–303. <https://doi.org/10.1111/1475-6773.12090>

- Fuller, B., Kagan, S. L., Loeb, S., & Chang, Y. (2004). Child care quality: centers and home settings that serve poor families. *Early Childhood Research Quarterly*, 19(4), 505–527. <https://doi.org/10.1016/j.ecresq.2004.10.006>
- Gelman, A. (2007). Struggles with survey weighting and regression modeling (with discussion and rejoinder). *Statistical Science*, 22(2), 153–164.
- Goldstein, H., & Healy, M. J. R. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, 158, 175–177.
- Grau, E., Potter, F., Williams, S., & Diaz-Tena, N. (2006). Nonresponse adjustment using logistic regression: To weight or not to weight? *Paper presented at the Joint Statistical Meetings*, Seattle, WA. Retrieved September 6, 2012 from <https://www.asasrms.org/Proceedings/y2006/Files/JSM2006-000717.pdf>
- Hahs-Vaughn, D. L., & Onwuegbuzie, A. J. (2006). Estimating and using propensity score analysis with complex samples. *The Journal of Experimental Education*, 75(1), 31–65. <https://doi.org/10.3200/JEXE.75.1.31-65>
- Hallberg, K., Steiner, P. M., & Cook, T. D. (2011). *The role of pretest and proxy-pretest measures of the outcome for removing selection bias in observational studies*. Presentation at the Spring 2011 Society for Research on Educational Effectiveness Conference. Retrieved September 1, 2012, from <https://www.sree.org/conferences/2011/program/downloads/abstracts/167.pdf>
- Hansen, M. H., Madow, W. G., & Tepping, B. J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78(384), 776–793.
- Holt, D., Smith, T., & Winter, P. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society, Series A*, 143, 474–487.
- Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, 101(475), 901–910.
- Hornik, R., Maklan, D., Judkins, D., Cadell, D., Yanovitzky, I., Zador, P., Southwell, B., Mak, K., Das, B., Prado, A., Barmada, C., Jacobsohn, L., Morin, C., Steele, D., Baskin, R., & Zanutto, E. (2001). *Evaluation of the national youth anti-drug media campaign: Second semi-annual report of findings* (Research Report). Westat. Retrieved October 2018 from <https://www.ncjrs.gov/App/publications/abstract.aspx?ID=191050>
- Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings among experimentalists and observationalists in causal inference. *Journal of the Royal Statistical Society Series A*, 171, 481–502.
- Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86, 4–29.

- Kish, L., & Frankel, M. (1974). Inference from complex samples (with discussion). *Journal of the Royal Statistical Society, Ser. B*, 36, 1–37.
- Korenman, S., Abner, K. S., Kaestner, R., & Gordon, R. A. (2012). The child and adult care food program and the nutrition of preschoolers. *Early Childhood Research Quarterly*, <https://dx.doi.org/10.1016/j.ecresq.2012.07.007>
- Korn, E. L., & Graubard, B. I. (1995a). Analysis of large health surveys: Accounting for the sampling designs. *Journal of the Royal Statistical Society*, 158, 263–295.
- Korn, E. L., & Graubard, B. I. (1995b). Examples of differing weighted and unweighted estimates from a sample survey. *The American Statistician*, 49(3), 291–295. <https://doi.org/10.1080/00031305.1995.10476167>
- Kurth, T., Walker, A. M., Glynn, R. J., Chan, K. A., Gaziano, J. M., Berger, K., & Robins, J. M. (2006). Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology*, 163, 262–270.
- Ladd, H. F. (2012). Education and poverty: Confronting the evidence. *Journal of Policy Analysis and Management*, 31(2), 203–227.
- Lenis, D., Nguyen, T. Q., Dong, N., & Stuart, E. A. (2019). It's all about balance: Propensity score matching in the context of complex survey data. *Biostatistics*, 20(1), 147–163. <https://doi.org/10.1093/biostatistics/kxx063>
- Lipsey, M. W., Hofer, K. G., Dong, N., Farran, D. C., & Bilbrey, C. (2013). *Evaluation of the tennessee voluntary prekindergarten program: Kindergarten and first grade follow-up results from the randomized control design* (Research Report). Vanderbilt University, Peabody Research Institute.
- Little, R. J., & Vartivarian, S. (2003). On weighting the rates in non-response weights. *Statistics in Medicine*, 22, 1589–1599.
- Magnuson, K. A., Ruhm, C., & Wald, J. (2007). Does prekindergarten improve school preparation and performance? *Economics of Education Review*, 26, 33–51.
- 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.
- Potter, F., Grau, E., Williams, S., Diaz-Tena, N., & Carlson, B. L. (2006). *An application of propensity modeling: Comparing unweighted and weighted logistic regression models for nonresponse adjustments* [Paper presentation]. Joint Statistical Meetings, Seattle, WA, United States. Retrieved September 6, 2012, from <http://www.amstat.org/sections/srms/proceedings/y2006/Files/JSM2006-000801.pdf>
- Ridgeway, G., Kovalchik, S. A., Griffin, B. A., & Kabeto, M. U. (2015). Propensity score analysis with survey weighted data. *Journal of Causal Inference*, 3(2), 237–249. <https://doi.org/10.1515/jci-2014-0039>
- Ridgeway, G., McCaffrey, D. F., Morral, A., Burgette, L., & Griffin, B. A. (2012). Toolkit for weighting and analysis of nonequivalent groups: A tutorial for the

- twang package. Retrieved September 6, 2012, from <http://cran.r-project.org/web/packages/twang/vignettes/twang.pdf>
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.
- Schonlau, M., Soest, A. V., Kapteyn, A., Couper, M., & Winter, J. (2004). Adjusting for selection bias in web surveys using propensity scores: The case of the health and retirement study. In *Proceedings of the section on survey research methods*. Retrieved September 6, 2012, from <https://www.amstat.org/sections/srms/Proceedings/y2004/files/Jsm2004-000032.pdf>
- Schuurmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing equivalence of average bioavailability. *Journal of Pharmacokinetics and Bio-pharmaceutics*, 15, 657–680.
- Scott, A. J., & Holt, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77, 848–854.
- Steiner, P. M., & Wong, V. C. (2018). Assessing correspondence between experimental and non-experimental estimates in within-study comparisons. *Evaluation Review*, 42(2), 214–247.
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1), 1–21.
- Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *The Journal of the Royal Statistical Society, Series A*, 174(2), 369–386.
- Tourangeau, K., Nord, C., Le, T., Sorongon, A. G., & Najarian, M. (2009). *Early childhood longitudinal study, kindergarten class of 1998-99 (ECLS-K)*, Combined User's Manual for the ECLS-K Eight-Grade and K-8 Full Sample Data Files and Electronic Codebooks (NCES 2009-004).
- Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, 6(4), 371–386.
- Tryon, W. W., & Lewis, C. (2008). An inferential confidence interval method of establishing statistical equivalence that corrects Tryon's (2001) reduction factor. *Psychological Methods*, 13(3), 272–277.
- U.S. Department of Education, National Center for Education Statistics (2009). *Early childhood longitudinal study, kindergarten class of 1998-99 (ECLS-K) kindergarten through eighth grade full sample public-use data and documentation* (NCES 2009005).
- Wilde, E. T., & Hollister, R. (2007). How close is close enough? Evaluating propensity score matching using data from a class size reduction experiment. *Journal of Policy Analysis and Management*, 26, 455–477.

- Wong, V. C., & Steiner, P. M. (2018). Designs of empirical evaluations of non-experimental methods in field settings. *Evaluation Review*, 42(2), 176–213. <https://doi.org/10.1177/0193841X18778918>
- Zanutto, E. L. (2006). A comparison of propensity score and linear regression analysis of complex survey data. *Journal of Data Science*, 4, 67–91.
- Zanutto, E. L., Lu, B., & Hornik, R. (2005). Using propensity score subclassification for multiple treatment doses to evaluate a national anti-drug media campaign. *Journal of Educational and Behavioral Statistics*, 30, 59–73.

Author Biographies

Nianbo Dong is an associate professor in the School of Education at the University of North Carolina at Chapel Hill. His primary research interests include causal inference, statistical power analysis, multilevel modeling, and program and policy evaluation.

Elizabeth A. Stuart is an associate dean for research and professor in the Department of Mental Health, the Department of Biostatistics, and the Department of Health Policy and Management of the Johns Hopkins Bloomberg School of Public Health. Her primary research interests are in statistical methodology for mental health and education research, particularly relating to causal inference and missing data and in methods for estimating population treatment effects.

David Lenis is a quantitative scientist at Covera Health. His primary research interests are in causal inference, particularly in the estimation of population treatment effects using nonexperimental data.

Trang Quynh Nguyen is an assistant scientist in the Department of Mental Health, Johns Hopkins Bloomberg School of Public Health. Her primary research interests are in statistical methods for causal inference with applications in public health.