A Comparison of Bias Reduction Methods:Clustering versus Propensity Score Subclassification and Weighting

A comparison of Bias Reduction Methods: Clustering versus Propensity Score Subclassification and Weighting

AUTHORS

Ida D'Attoma, Ph.D. Senior Assistant Professor in Statistics for Economics at Department of Statistical Sciences, University of Bologna, via Belle Arti,41 -40126- Bologna (Italy), phone:0039(051)2094613, E-mail: Ida.dattoma2@unibo.it

and

Furio Camillo, Associate Professor of Business Statistics and Data Mining at Department of Statistical Sciences, University of Bologna, via Belle Arti,41 -40126- Bologna (Italy), phone:0039(051)2098251, E-mail: furio.camillo@unibo.it

and

M. H. Clark, Ph.D. Associate Lecturer at Department of Educational and Human Sciences, College of Education and Human Performance, P.O. Box 161250 University of Central Florida, Orlando, FL 32816-1259, phone: 407-823-0442, E-mail: mhclark@ucf.edu

Abstract

Propensity score (PS) adjustments have become popular methods used to improve estimates of treatment effects in quasi-experiments. While researchers continue to develop PS methods, other procedures can also be effective in reducing selection bias. One of these uses clustering to create balanced groups. However, the success of this new method depends on its efficacy compared to that of the existing methods. Therefore, this comparative study used experimental and non-experimental data to examine bias reduction, case retention, and covariate balance in the clustering method, PS subclassification, and PS weighting. In general, results suggest that the cluster-based methods reduced at least as much bias as the PS methods. Under certain conditions, the PS methods reduced more bias than the cluster-based method and under other conditions the cluster-based methods were more advantageous. While all methods were equally effective in retaining cases and balancing covariates, other data-specific conditions may likely favor the use of a cluster-based approach.

Introduction

Selection bias is one of the most vexing challenges to program evaluation. Problems associated with selection bias arise when one aims at assessing a policy's effectiveness with poorly matched comparison groups. Not all old or new solutions are equally effective in reducing selection bias and in producing well-matched cases. Traditionally, researchers have attempted to reduce potential selection bias by accounting for covariates in multiple regressions and analyses of covariance (ANCOVAs). However, these approaches are not appropriate considering that the purpose of

traditional covariate adjustment is to account for the shared variance between the covariates and the dependent variable, not balance the covariates between groups. While adding linear covariates to causal models may reduce the error variance, these models assume that the relationships between covariates and the dependent variable are the same in the treatment and control groups. However, if the distributions of covariates in each group are different (as is often the case in nonrandomized studies) is it unlikely that their relationships with the dependent variable are similar (Schafer & Kang, 2008). To effectively balance covariates, selection bias must be modeled by considering how covariates relate to selection into groups. In recent years, propensity score methods have become popular approaches for creating and applying these models to improve causal estimates. Propensity scores are the predicted probabilities that participants will select a treatment, given a set of observed covariates that influence their selection (Rosenbaum & Rubin, 1984). In theory, propensity scores should allow researchers to simultaneously balance all observed covariates between treatment and comparison groups, which will reduce selection bias. This balance is most commonly achieved by matching, weighting or subclassifying participants on the propensity score or including the propensity score as a covariate in a regression or ANCOVA. Over the last three decades, PS adjustments have gained popularity in many fields: Program evaluation (e.g., Harknett, 2006; Peck, 2007), economics (e.g., Dehejia & Wahba, 1999; Gerfin & Lechner, 2002; Imbens, 2004; Smith & Todd, 2005), political science (e.g. Ho, Imai, King, & Stuart, 2007), sociology (e.g., Morgan & Harding, 2006), medicine (e.g. Christakis & Iwashyna, 2003), and educational research (e.g., Hong & Raudenbush, 2005; Tebes et al., 2007). While specific propensity score methods may vary in their ability to reduce bias in individual studies, several researchers have shown that, in general, propensity scores can substantially reduce bias in observed covariates (Austin, 2014; Bai, 2013; Garrido, et al., 2014; Stone & Tang, 2013).

Although other comparative studies that examined propensity score adjustments do not consistently agree on the best method, PS subclassification tends to be one of the more effective ones. Schafer and Kang (2008) found that PS subclassification reduced more bias than either

PS matching or weighting, but was comparable to traditional covariate adjustment. When Zanutto (2006) incorporated survey weights in her comparison of PS subclassification with a traditional covariate adjustment, she also found that the two methods were similar in their ability to reduce selection bias and balance and covariates. Shadish et al. (2008) compared traditional covariate adjustments to several PS adjustment methods (PS subclassification, linear PSs in an ANCOVA, non-linear PSs in an ANCOVA, PS weighting) on two educational outcomes and found varying results. For one outcome variable, PS subclassification reduced more bias than the other PS adjustments, but was not more effective than the traditional covariate adjustment. However, the most effective adjustment method was a doubly robust procedure that combined the two methods to estimate treatment effects from propensity score subclasses while accounting for the same variables used to estimate the propensity scores as individual covariates. For the second outcome variable, PS weighting was the most effective PS adjustment method, and accounting for individual covariates in additional to the PS adjustments did not improve treatment estimates. In some studies, PS subclassification did not reduce the most bias, but it balanced individual covariates better than other methods (Austin & Mamdani, 2006; Harder et al., 2010). Although several researchers have found that PS weighting can be problematic (Freedman & Berk, 2008; Rubin, 2001), others have found that it can remove as much as 95% to 100% of bias if the propensity score model is correctly specified (Hong, 2007, p.16; Kang & Schafer, 2007, p. 534; Lunceford & Davidian, 2004, p. 21) .Building on previous research that compared various methods for reducing selection bias, we expect that this comparative study between the cluster-based approach and PS subclassification and PS weighting will provide methodological guidance to analysts and researchers who are similarly vexed by selection bias.

Within the last few years, another promising new method, which uses cluster analyses and a global measure of imbalance, has emerged for reducing selection bias in nonrandomized experiments. Although cluster analysis (CA) has a long history of use in program evaluation, it is more

commonly accepted and applied in experimental studies (e.g., Gibson, 2003; Yoshikawa, Rosman & Hsueh, 2001). Peck (2005) used CA to identity subgroups within experimental data to understand variation in program impacts. Because CA is formulated using only baseline characteristics, which are, by design, exogenous to the treatment, she was able to successfully estimate unbiased treatment effects for the program she was evaluating. The new-cluster based method proposed here has two unique features (Camillo & D'Attoma, 2012; Peck at al. 2012). First, the use of cluster analysis increases the possibility of finding local spaces in which variables are balanced without the typical problems related to the estimation of probability models (e.g., functional forms, minimum sample sizes). In cluster analysis, for example, there is no general rule of thumb regarding minimum sample sizes or the relationship between units and the number of clustering variables. CA will always render a result irrespective of how many variables are used or how small the sample size is (e.g., Mooi & Sarstedt, 2011; Tonks, 2009). The second feature, the measure of global imbalance (GI), is a single measure of the imbalance in the data based on the concept of inertia[1] as a measure of dependence between a set of pre-treatment covariates and the treatment assignment indicator. While the common practice is to assess the extent to which comparison and treatment groups are matched on each variable, the GI measure provides an overall assessment of how well cases are matched. The GI measure differentiates the cluster-based method from the conventional use of cluster analysis, as used by Peck (2005), where balance is not assessed.

Using two simulated samples, one with a binary treatment case and another with a multi-treatment case, Camillo & D'Attoma (2012) compared the performance of the GI measure used in a subgroup analysis to (a) the *L1* distance measure of Iacus et al. (2011), (b) the *In* test statistic

[1] Here, the term inertia is used in analogy with applied mathematics, where the "moment of inertia" is the integral of mass times the squared distance to the centroid (Greenacre, 1984).

of Li et al. (2009), and (c) the Hansen and Bowers' test. Not only did the results suggest that the GI measure outperformed the other methods, but that it also reached the maximum percent bias reduction in most of the examined clusters. These results encouraged us to compare its performance to other well-established adjustment methods using real data.

While there are several widely accepted conventional adjustments to selection bias, this paper examines three approaches to overcome selection bias: the new cluster-based method (Peck et al. 2010; D'Attoma & Camillo, 2011) , the traditional propensity score (PS) subclassification method (Rosenbaum & Rubin, 1984) and the inverse probability of treatment weighting (IPTW) method (Horvitz & Thompson, 1952; Austin & Stuart, 2015). PS subclassification (also referred to as PS stratification) and IPTW are among the most common propensity score methods used in applied academic research. Due to their popularity and demonstrated effectiveness in reducing selection bias, they serves as good methods by which to compare other bias-reduction methods. Therefore, the objective of this study is to ascertain if the cluster-based method is more effective than well-established propensity score methods in reducing selection bias and retaining cases.

Methods

Data

The dataset from the work of Shadish et al. (2008) was used for this study, because it provided a clear benchmark for assessing selection bias and consisted of real-world, individual observations, thereby reducing the weaknesses of other approaches, such as, computer simulations, single-study comparisons and meta-analysis. As a way of measuring how well nonrandomized experiments can approximate treatment effects found in randomized experiments, participants were randomly assigned to either a randomized experiment or a nonrandomized experiment. In the randomized experiment, the participants were randomly assigned to mathematics or vocabulary training; in the nonrandomized experiment, the

participants chose their training. Several variables expected to predict the condition that participants chose were measured using a battery of pretests. These included demographic characteristics: Age, education, marital status, and major area of study; educational assessments: ACT or SAT scores, college and high school grade point averages (GPA) and vocabulary and mathematical tests; and personality traits: Extroversion, emotional stability, agreeableness, openness to new experiences, conscientiousness, mathematics anxiety, and depression. Table 1 illustrates a complete list of covariates and how they relate to treatment selection and outcomes for both the randomized and nonrandomized samples.

----------INSERT TABLE 1 ABOUT HERE----------Of the 445 undergraduate students from introductory psychology classes who volunteered to participate in this study, 235 were randomly assigned to the randomized experiment and 210 to the nonrandomized experiment. Of the 235 participants in the randomized experiment, 116 were randomly assigned to vocabulary training and 119 to mathematics training. Of the 210 participants in the nonrandomized experiment, 79 selected mathematics training and 131 selected vocabulary training.

After receiving their respective training, all the participants were measured on two outcomes: vocabulary and mathematics. The 50-item post-test contained 30 vocabulary items (15 items presented in training and 15 new items) and 20 mathematics items (10 presented in training and 10 new items). For the math posttest, those who received the mathematics training served as treatment group and those in the vocabulary training served as the control group. For the vocabulary posttest, those who received the vocabulary training served as treatment group and those in the mathematics training group served as the control group. Therefore, this design created four experiments: a randomized experiment to test a vocabulary intervention, a randomized experiment to test a mathematics intervention, a nonrandomized experiment to test a vocabulary intervention,

and a nonrandomized experiment to test a mathematics intervention. Since the data collection and treatment procedures for each training method were identical in all respects except in the assignment method, it is reasonable to assume that the only difference in these outcomes were due to the assignment method. Therefore, selection bias for each intervention could be easily measured by comparing the treatment effects from the nonrandomized experiment to the treatment effects from the randomized experiment.

Statistical Analysis

PS subclassification, IPTW, and the cluster-based method were compared to assess the extent by which they can reduce estimated bias using the results obtained from a randomized experiment as a benchmark. Since propensity score methods may be less effective in reducing bias if the propensity score models are misspecified (Kang & Schafer, 2007; Lunceford & Davidian, 2004), we also included doubly robust models for PS subclassification and IPTW.

Using Propensity Scores Subclassification. For unit i ($i = 1, \ldots, N$), let T_i denote the treatment assignment (in this case $T_i = 1$ if the unit i receives the vocabulary training, and $T_i = 0$ if the unit receives the mathematics training) and let $\underline{x}_i$ denote the vector of observed covariates. The propensity score ($e(\underline{x}_i)$) for unit i, as proposed by Rosenbaum and Rubin (1983), is defined as the conditional probability that the unit i receives the treatment, given the set of observed covariates,

$$e(\underline{x}_i) = \Pr(T_i = 1 | X_i = x_i) \qquad (1)$$

Rosenbaum and Rubin (1983) have shown that the methods, which balance groups on $e(X)$, produce unbiased estimates of the treatment effects, if the strong ignorability assumption holds, that is treatment assignment (T) and the potential outcomes [$Y(0), Y(1)$] are conditionally independent given the observed covariates as in eq. (2)

(2)

and if eq. (3) holds.

$$0 < P(e(x_i)) < 1 \qquad (3)$$

Once the propensity scores are estimated, units are grouped based on their estimated propensity scores. Although propensity scores may be grouped in various ways, Cochran (1968) and Rosenbaum and Rubin (1984) found that one may expect approximately 90% reduction in bias when balancing units by quintiles. Following these recommendations, we estimated propensity scores using a logistic regression in which all covariates were used as predictors to estimate the predicted probability that each unit would select the vocabulary intervention. Then, units in both treatment groups were classified into quintiles based on their propensity scores. For each variable used as covariate, the balance achieved by subclassification was examined using a standardized measure of bias similar to that used by Rosenbaum and Rubin (1985). Odds ratios were computed for the categorical covariates, then converted to standardized estimates of bias using a formula similar to Hasselblad and Hedges's (1995). The standardized bias estimates and equation used to compute them are provided in Appendix A.

Using Inverse Probability of Treatment Weighting. With IPTW each unit is weighted by the inverse of the probability of receiving the treatment that the unit actually received. In particular, treated individuals receive an IPTW equal to (4):

$$1/e(x_i) \qquad (4)$$

and control individuals receive a weight equal to (5) (e.g., Harder et al., 2010; Lanehart et al., 2012).

$$1/\big(1 - e(x_i)\big) \qquad (5)$$

The weights are then used in a weighted least squares regression model along with other predictor covariates. The main characteristic of IPTW is that it includes all subject in a study, causing no loss of sample as in other methods. The main disadvantage of using this conditioning method is the potential bias caused by the possible presence of extreme propensity scores (e.g., Lanehart et al., 2012; Austin, 2011; Shadish & Steiner, 2010). In the presence of extreme propensity scores, stabilization techniques can be used to attenuate bias (e.g. Harder et al. 2010; Robins et al. 2000).

Using Cluster Based Method. The general idea underlying the cluster-based method is to balance unequal group distributions by stratifying on clusters based on several pre-treatment covariates and using a global measure of imbalance (Peck et al. 2012; Camillo & D'Attoma, 2012; D'Attoma & Camillo, 2011; Peck, 2005). Cluster analysis does not create a single aggregate score, permitting each covariate to maintain its functional form. In general, balancing by clusters, rather than by units, increases common support, which is the likelihood that each unit in the treatment group will be compared to a similar unit in the control group. Although, this increased common support often means that units in the clustered analysis are less similar to each other than when matching pairs of units, group homogeneity can be controlled by varying the number of clusters. As with other subclassification adjustments, fewer clusters retain more participants and more clusters allow for better covariate balance. Finding the number of clusters will be illustrated and explained further in the results section.

The global measure of imbalance is used to evaluate balance within each cluster solution. As explained in Camillo & D'Attoma (2010), it originates from the decomposition of the variability (total inertia) associated with a matrix into the variability due to covariates (X) (*inertia within*) and the variability caused by the selection into treatment mechanism (T) (*inertia between*), which is called GI (Camillo & D'Attoma, 2010, p. 179).

The GI measure can be computed according to the formula in eq. (6):

$$GI = \frac{1}{Q}\sum_{t=1}^{T}\sum_{j=1}^{J_Q}\frac{b_{tj}^2}{k_{.t}k_{.j}} - 1 \qquad (6)$$

where Q denotes the number of pre-treatment covariates, T denotes the number of treatment levels, J_Q denotes the set of all categories of the Q baseline variables, b_{tj} is the number of units in category j in treatment group t. $k_{.t}$ is the treatment group size and $k_{.j}$ is the number of units in category j. According to the GI measure, perfect imbalance occurs when the between-groups inertia $[I_B]$ equals the total inertia $[I_T]$, which implies that the observed total variability of the covariates X-space is completely due to the selection-into-treatment mechanism. Thus, the proposed GI measure in data varies in $[0, I_T]$. Then, following D'Attoma & Camillo (2011) and Peck et al. (2012), we determine the significance of the GI measure by performing an imbalance test of the null hypothesis of independence between X and T. According to such a test (for more details, see D'Attoma & Camillo, 2011; Camillo & D'Attoma, 2012), the interval of plausible values for GI is defined as in (7):

$$GI \in \left(0, \frac{\chi^2_{(T-1)(J-1),\alpha}}{nQ}\right) \qquad (7)$$

If the GI calculated on the specific dataset is outside the interval, then the null hypothesis is rejected and data are deemed unbalanced. Estadella et al. (2005) have demonstrated that the length of the interval depends on *n*, *Q*, *J* and *T*. As a consequence, its length decreases with sample size *n*. When *n* increases, the length decreases and it becomes more plausible to reject the null and deem data unbalanced. For more comprehensive explanations of the theoretical framework of the GI measure, its related test, and illustrations of its use and effectiveness, see D'Attoma & Camillo (2011), Camillo & D'Attoma (2012), and Peck et al. (2012).

Although some might argue that significance testing is inadequate to evaluate covariate balance because it is affected by the sample size, this test does not depend only on *n*. To avoid potential critiques and improve reliability of results we have opted for an additional statistic to evaluate global imbalance in data, that is the multivariate imbalance coefficient (MIC) proposed in D'Attoma & Camillo (2011). The MIC is an index that ranges between 0 and 1 and it is defined as one minus the ratio between the within-groups inertia relative to the total inertia as in (8):

$$MIC = 1 - \frac{I_W}{I_T} \qquad (8)$$

MIC=0 denotes perfect balance; whereas, MIC=1 indicates perfect imbalance. It expresses the importance of the measured imbalance relative to the total inertia.

In short, the method is carried out in three steps for estimating treatment effects in non-experimental data in which selection bias on observed variables is minimized without drawing upon the specification and the estimation of a probabilistic model.

Step 1: Measure and test balance. Compute the MIC index, the GI measure on the overall unclustered data and test its statistical significance.

Step 2: If imbalance exists, which is probable in all non-experimental data, perform a subgroup analysis, which involves implementation of cluster analysis to identify homogeneous groups. Based on previous studies (Green & Krieger, 1995; Lebart, Morineau, & Piron, 1997; Peck et al., 2012), we suggest using Ward's algorithm on multiple correspondence analysis (MCA) factorial coordinates, where the proximity between two groups is measured by the squared Euclidean distance. By using MCA as an intermediate step before cluster analysis, it is possible to exploit the advantage of working with continuous variables (the MCA factorial coordinates) rather than categorical covariates (some or all the original variables), which need to be treated with unusual metrics.

Step 3: Measure and test the balance in each cluster and compute local treatment effects in balanced clusters, and prune the observations in unbalanced clusters.

All the three steps may be easily carried out using the %GI SAS macro (Camillo & D'Attoma, 2012) that can be downloaded at https://www.jstatsoft.org/article/view/v051c01

Results

The Effects of Mathematics Training on Mathematics Outcome: The Benchmark and the Propensity Score Results

The effects estimated from randomized experiments are considered the best against which all adjusted and unadjusted nonrandomized results are compared (Shadish et al. 2008). Since these data include observations from both randomized and nonrandomized experiments, the randomized experiments made it possible to measure the selection bias in the nonrandomized studies. Because some of the covariates in the randomized experiments were not balanced due to sampling error, the average treatment effects (ATEs) were adjusted following the procedures recommend by Schafer and Kang (2008), in which each outcome was estimated from the group differences in the randomized experiment after accounting for all of the available covariates and interactions between these covariates and treatment conditions that were statistically significant at $p < .1$. This covariate-adjusted effect from the randomized experiment (reported in Table 2) was the benchmark against which the performances of the PS subclassification, IPTW, and the cluster-based methods were compared. In particular, the participants who received mathematics training performed better on the mathematics outcome by 3.98 points than those who received vocabulary training. The absolute value of the difference between the

randomized covariate-adjusted results and those of the unadjusted nonrandomized experiment (Δ=|3.98-5.01|=1.03) is a measure of the bias in the unadjusted nonrandomized results.

INSERT TABLE 2 ABOUT HERE

After making some adjustments to expression (2.6) in Rubin (1973), the percent bias reduction (p.bias) was calculated as in (9):

$$p.bias = \left[\frac{|b_{y(x)}^{U} - b_{y(x)}^{A}|}{|b_{y(x)}^{U}|}\right] \times 1 \qquad (9)$$

where $\bar{y}^{U}$ and $\bar{y}^{A}$ are respectively the unadjusted and adjusted estimates of the average treatment effect.

The propensity score subclassification on propensity score quintiles (Rosenbaum & Rubin, 1984), already implemented by Shadish et al. (2008), was replicated. Like the previous study, propensity scores were created using logistic regression and included the following covariates: Vocabulary pretest, mathematics pretest, log(number of prior mathematics courses), like mathematics, preferring literature over mathematics, extraversion, conscientiousness, emotionality, mathematics anxiety, log(depression), Caucasian, male, mother's education, father's education, math-intensive major, college GPA, mother's education squared, and college GPA squared. All five subclasses were balanced according to the conventional variable-by-variable tests, and no units were discarded. The percent of bias reduction was similar to that estimated by Shadish et al. (2008), wherein ATE=3.72, *SE*=.57 and p.bias=75%.

Using the traditional PS subclassification approach, Shadish et al. (2008) were able to reduce selection bias by 71% and by 96% by including additional covariates in the PS model, thereby reducing the difference between the randomized and adjusted nonrandomized effects to nearly zero. Using slightly different randomized treatment effects, we were able to reduce bias by 75% and 93%, respectively. Due to the additional bias reduced by the doubly robust approach, it is likely that the propensity scores were not modeled correctly, despite testing several interactions and including non-linear terms. Therefore, our comparisons with other models will focus on the results from the PS subclassification with additional covariates.

INSERT TABLE 3 ABOUT HERE

The Effects of Mathematics Training on Mathematics Outcome: The IPTW results

Using the same propensity scores estimated for PS subclassification, weights were computed according to formula (4) and (5). In order to adjust for potential bias due to extreme weights, following Lanehart et al. (2012), normalized weights were computed. The normalized weights were then used to estimate the desired treatment effect via weighted least squares regression. Table 3 shows that the IPTW method alone reduced the bias by 63%, and IPTW with additional covariates reduced bias by 75%. Given this difference between the two methods, we will use the second model as the basis for comparison to the other methods.

The Effects of Mathematics Training on Mathematics Outcome: The Cluster-based Method Results

The extent by which similar or better bias reduction can be obtained by using the cluster-based method was examined. The same covariates used to estimate propensity scores in the previous adjustments were used for the cluster analysis. The three-step cluster-based method was performed as detailed below:

Step 1: The GI measure for the overall data was computed using the %GI SAS macro (Camillo & D'Attoma, 2012). The resulting value of 0.03646 (shown in Table 4) can be interpreted as a reflection of an imbalance in the data. The GI measure falls in the critical region, thereby requiring a statistical adjustment to estimate an unbiased treatment effect. Unlike PS methods, this method provides a single measure of sample balance that, by construction, takes into account all possible interactions between variables.

INSERT TABLE 4 ABOUT HERE

Step 2: The second step involves cluster analysis to identify homogeneous groups on the basis of MCA coordinates (Benzecri, 1973). The cluster analysis was carried out in Spad v6 software employing Ward's algorithm on the MCA coordinates, where the proximity between two groups is the square of the Euclidean distance between them. The MCA was carried out using the following variables: Vocabulary pretest, mathematics pretest, log (number of prior mathematics courses), like mathematics, preferring literature over mathematics, extraversion, conscientiousness, emotionality, mathematics anxiety, log(depression), Caucasian, male, mother's education, father's education, math-intensive major, and college GPA. Twelve of these variables were originally continuous, but grouped in classes prior to including them in the MCA analysis. The result of the MCA is a set of new variables (factorial coordinates) that are continuous and orthogonal to one other.

The MCA coordinates obtained were used as input for cluster analysis. Although cluster analysis is an atheoretical approach, the covariates we included in the preceding MCA were known predictors of treatments and selection.

Figure 1 is a dendrogram (or tree diagram) used to document the clustering process and to choose the number of clusters. The dendrogram produces sequences of nested partitions of increasing heterogeneity between a partition into K clusters, where each unit is isolated, and a partition into one cluster which includes all the units. The dendrogram can be cut at a greater or lesser height to obtain a smaller or larger number of clusters; this number can be chosen by the statisticians or practitioners to optimize certain quality criteria. The main criterion is the loss of between-cluster sum of squares represented in Figure 1 by the height of the pairs of connecting branches. Since this loss must be as small as possible, the dendrogram should be cut at the level where the height of the branch is large. Such large gaps between clusters arguably indicate natural groups. Unfortunately, natural clusters are not always obvious and may differ according to the algorithm used to compute them. In practice, the determination of the "real" number of clusters must be empirical as much as theoretical. From a practitioner's point of view, cluster solutions must be clearly evident rather than theoretically optimal. In an evaluation context, determining natural clusters is easier because both covariate balance and common support must be considered, which limits the number of acceptable cluster solutions, making optimal choices more obvious. The purpose of this comparative study is not to find one optimal solution, but to explore all possible solutions from the dendrogram to study how the percent of bias reduction relates to the number of clusters. It was decided to examine 18 partitions, from 2-cluster partition to 20-cluster partition[2].

INSERT FIGURE 1 ABOUT HERE

[2] We limited the analysis to a 20-cluster partition because a partition with more than 20 clusters might not have sense from a practitioner's point of view.

GI was measured and tested for each of the 18 cluster solutions using three different alpha levels. As reported in Table 5, the smallest cluster solutions ($k < 9$) did not satisfy the balance criterion; and the largest cluster solutions ($k > 16$) did not satisfy the common support criterion. However, cluster solutions 9 through 16 satisfied both balance and the common support criteria. Therefore, any of these solutions are acceptable.

INSERT TABLE 5 ABOUT HERE

As can be seen from Table 5 and Figure 2, when there is common support, more clusters result in better group balance and fewer discarded units. As the number of groups increases, cluster size and the probability of common support decrease. The red vertical line in Figure 2 indicates the point in which the common support ceases to hold. In this case, common support declines with more than 16 clusters.

INSERT FIGURE 2 ABOUT HERE

Step 3: During the final stage of the procedure, the effects of interest were analyzed and the percent bias of reduction was examined. Specifically, the treatment effect was measured in all eight cluster solutions that satisfied both balance and common support criteria. Clearly, the more clusters

we retain, the smaller the clusters sizes will be, and statistical power might diminish especially in studies like this that have so few observations. For all eight cluster partitions the effect of mathematics training on the mathematics outcome was computed with the related standard errors[3].

The weighted sum of the average treatment effect for each k-cluster partition (ATE_d) was estimated by (10):

$$ATE_d = \sum_{k=1}^{K} \frac{n_k}{N} ATE_k \qquad (10)$$

where *k* is the number of clusters in the partition, n_k is the number of participants in each cluster, N is the total sample size of the study, and ATE_k is the average treatment effect for cluster *k*. Table 6 reports the weighted treatment effect with standard errors in parentheses, the absolute bias and the percent bias reduction for each K-cluster partition for the mathematics outcome. Absolute bias and percent bias reduction were computed using the ATE from the covariate-adjusted randomized experiment as the benchmark for comparison.

INSERT TABLE 6 ABOUT HERE

The average treatment effect of all the selected partitions (ATE_{TOT}) was computed by (11):

$$ATE_{TOT} = \frac{1}{D} \sum_{d=1}^{D} ATE_d \qquad (11)$$

[3] The standard errors are computer using the formula in Zanutto (2006, p. 70)

where *D* is the number of considered partitions and *d* is the specific partition ($d \in D$). The computation of the ATE of all the acceptable cluster solutions, avoids the choice of the best cluster partition that might introduce more subjectivity in the analysis. However, ATE_{TOT} might reduce less bias than some single cluster solutions. For the mathematics intervention, the average treatment effect (ATE_{TOT}) of all the selected partitions was 4.16, the average absolute mean difference (Δ) was 0.18 and the average percent bias reduction was 82.52%

INSERT FIGURE 3 ABOUT HERE

It clearly emerges from Figure 3 and Figure 4 that the selection of a k-cluster partition with a greater number of clusters leads to increased percent bias reduction and produces more precise estimates. If one prefers to select the single best partition instead of using the average treatment effect of all the selected partitions, the best option in this case is the 15-cluster partition because it reduces the most bias (97%). Within the 13,15,16-cluster solutions, Tables 7,8 and 9 respectively, show that the average treatment effects are heterogeneous. However, computing the average effect of all k-cluster solutions might obscure this heterogeneity. The ability to examine heterogeneity in the effects is another advantage of subgroup analysis that conventional PS subclassification analyses may lack since subclasses are not "natural" groups[4]. Instead, PS bins and their number are based on an artificial division of data.

[4] For an in-depth analysis of unsupervised learning techniques to identify natural group structures underlying the data see for example Everitt (1993; 2011); Duda et al. (2001) and Hastie et al. (2001).

This is due to the different essence of cluster analysis. Cluster analysis is a technique that "finds natural groups by sorting the observations into groups such that the degree of natural association is high among members of the same group and low between members of different groups" (Anderberg, 1973, p.3). The definition of natural group is not obvious, but it is typical of data mining and natural language processing that provide different methods to find patterns. A group is natural when it already exists in the unstructured data[5] and the algorithms detect it on the basis of similarity of objects within it. As reported in Anderberg (1973, p.5), "since similar observation are grouped together, the individuals tend to assume class labels and the whole process may give names to things"

INSERT TABLE 7 ABOUT HERE

INSERT TABLE 8 ABOUT HERE

INSERT TABLE 9 ABOUT HERE

The Effects of Vocabulary Training on Vocabulary Outcome: The Benchmark and the Propensity Score Results

5 Unstructured data do not have an underlying pre-defined data model as is not organized in a pre-defined manner.

Following the procedures used for the mathematics outcome, the covariate-adjusted effect from the randomized experiment (reported in Table 10) became the benchmark against which the performances of the PS subclassification and the cluster-based method were compared. In particular, the participants who received vocabulary training performed better on the vocabulary outcome by 8.25 points than those who received mathematics training. The absolute value of the difference between the randomized covariate-adjusted results and those of the unadjusted nonrandomized experiment ($\Delta=|8.25-9.00|=.75$) is the measure of the bias in the unadjusted quasi-experiment results from the vocabulary outcomes.

INSERT TABLE 10 ABOUT HERE

Using the traditional PS subclassification approach, Shadish et al. (2008) were able to reduce selection bias by 87%. Although including additional covariates in the PS subclassification model did not reduce much more bias than the model without the added covariates, it did make a small improvement (see Table 11). Since it is not clear from this difference whether or not propensity scores had been modeled correctly, we will compare the results from both PS subclassification methods to IPTW and cluster methods.

INSERT TABLE 11 ABOUT HERE

The Effects of Vocabulary Training on Vocabulary Outcome: The IPTW results

Following the procedures for the mathematics outcomes, the propensity scores computed for the other PS methods were weighted according to formula (4) and (5) using normalized weights, and the normalized weights were used to estimate treatment effect. Table 11 shows that the IPTW method alone reduced bias by 96%, and the IPTW with additional covariates reduced bias by 91%.

The Effects of Vocabulary Training on Vocabulary Outcome: The Cluster-based Method Results

The extent by which similar or better bias reduction could be obtained by using the three-step cluster-based method described previously was examined. Steps 1 and 2 from the mathematics outcome were replicated for this outcome. However, in the third step, the effects of vocabulary training on vocabulary outcome were analyzed and the percent bias reduction examined. As was done for the mathematics outcome, the treatment effect was measured in all eight cluster partitions that satisfied both balance and common support criteria. Table 12 reports the ATE_d with standard errors in parentheses, the absolute bias, and the percent bias reduction relative to each selected K-cluster partition.

INSERT TABLE 12 ABOUT HERE

The average treatment effect (ATE_{TOT}) of all the selected partitions was 8.07, the average absolute mean difference (Δ) was 0.18 and the average percent bias reduction was 76%. For the vocabulary outcome, the 16-cluster partition was most effective in reducing bias with a percent bias reduction of 97%. Within the 16-cluster solution, Table 13 shows the heterogeneity of the average treatment effects for each cluster.

INSERT TABLE 13 ABOUT HERE

Propensity Score Methods versus the Cluster-based Method

When comparing the average reduction in bias for all eight k-cluster solutions that met the balance and common support standards to the PS subclassification and IPTW methods, the cluster-based method reduced less bias (82.52%) in the mathematics outcome than the PS subclassification method with added covariates (93%), but was still better than the IPTW with added covariates (75%). However, the 15-cluster partition reduced bias by 97%, which was better than either of the PS methods. For the vocabulary outcome, the average bias reduction from the eight k-cluster solutions (76%) was less than that from either PS subclassification method (91% with added covariates and 87% without added covariates) and either IPTW (96% without added covariates and 91% with added covariates). However, using only the 16-cluster partition with the cluster-based method reduced *more* bias (97%) than any of the PS methods.

The standard error bar graphs depicted in Figure 4 and Figure 5 show no relevant difference in the two methods with respect to the precision of the treatment effects, but generally, the estimates produced by the cluster-based method are, for some cluster solutions, slightly more precise than estimates produced by PS subclassification or by IPTW. Though several factors can affect the precision of an estimate, sample size is always a factor because the standard error formula includes the sample size in the denominator. Hence, with all other factors held steady, as sample size increases, the standard error decreases and the treatment estimates are more precise. For such a reason, it does not make sense to compare standard errors of PS adjustments with standard errors of ATE within a single cluster, since the sample size of a single cluster is always smaller than that

from the entire data set. Therefore, in Figure 4 and Figure 5, standard errors of PS subclassification are compared to the standard errors of the ATE_d calculated on the overall k-clusters partition.

INSERT FIGURE 4 ABOUT HERE

INSERT FIGURE 5 ABOUT HERE

Discussion and Conclusions

Bias Reduction

The aim of this paper was to examine the effectiveness of the cluster-based approach when compared to PS subclassification and IPTW in terms of percent bias reduction. In general, the propensity score and cluster-based methods were comparable in their ability to reduce selection bias. Under certain conditions, the PS methods reduced more bias than the cluster-based method and under other conditions the cluster-based methods were more advantageous. PS subclassification using a doubly robust analysis consistently reduced more bias than the total average treatment effect computed for all the eight k-cluster solutions. Whereas the IPTW reduced more bias than the total average treatment effect computed for all eight k-cluster solution only for the vocabulary outcome. However, specific cluster solutions (e.g., 15-clusters for the mathematics outcome and 16-clusters

for the vocabulary outcome) were more effective in reducing bias than the PS methods. Likewise, there is no significant difference in the precision of the estimates between the three methods. However, for the vocabulary outcome, standard errors from the cluster-method tended to be smaller (0.42-0.51) than those from PS subclassification (0.49-0.60) and IPTW (0.38-0.49).

Case Retention

In this particular study, we found that all of the methods retained the same number of units. However, this may not always be true for studies that use PS subclassification where (a) many subclasses are used and (b) the covariates are very strong predictors of selection, which reduces the common support between the propensity scores in the treatment and control groups. In such cases, the cluster-based method may allow for better case retention, which could improve external validity. On the other hand, some argue that the atheoretical nature of the cluster-based approach limits the ability of this method to produce generalizable findings (Peck et al. 2012). Since the IPTW method has neither of these limitations, it may be the best method with respect to generalizability.

Advantages to Using Cluster Analysis over Propensity Score Adjustments

While the amount of bias reduced by the cluster-based approach may not be significantly different than that from the PS subclassification method in *this* study, other studies have shown that the cluster-based method can be more effective in reducing bias than propensity score methods (Clark, 2011). Furthermore, the cluster-based approach offers some advantages over the propensity score methods. First, the cluster-based method balances unequal group distributions by stratifying on the covariates themselves, rather than on an aggregate of the covariates, permitting each covariate to maintain its functional form.

Second, it reduces the subjectivity of covariate selection that propensity score researchers are faced with. The choice of covariates and interaction terms or higher order terms to be included in statistical models plays a fundamental role in their ability to reduce selection bias. However, the criteria that propensity score researchers use to include covariates in propensity score models are inconsistent and many do not include any interactions or higher order terms (Thoemmes & Kim, 2011). Some researchers include any covariate that has a standardized bias less than 5 (e.g., Bai, 2013), others rely on significance tests (e.g., Diamond & Sekhon, 2013), and others include all available variables that are theoretically related to outcomes (e.g., Harknett, 2006). While there are algorithms that automatically include interactions and nonlinear covariates for propensity score estimations (e.g., Matchit in R), not all researchers regularly use these packages. As demonstrated in our own results, when propensity score models are misspecified, propensity scores alone may not sufficiently reduce bias. When using cluster analysis, researchers frequently choose covariates by applying CA to the observations' factor scores derived from a previous factor analysis. The implementation of CA on factor scores allows to use an abundance of clustering variables of different nature (qualitative or quantitative)[6] and with a different degree of correlation or association.

Another clear advantage of the cluster-based method is that it captures heterogeneity in the effects on natural groups, which cannot be done by matching or propensity score adjustments. Average treatment effects often obscure the changing effects within heterogeneous populations; the cluster-based approach capitalizes on this heterogeneity and allows for impact estimates within subgroups that might otherwise fail to be easily recognized. Particularly if subgroup effects operate in opposite directions, the overall treatment effect will be estimated as zero.

[6] Tonks (2009) provides a discussion of segment design and the choice of clustering variables in consumer markets.

Generally, clustering yields differentiated clusters that remain stable when small changes in data occur. As reported in Everitt (1993) and Jolliffe et al. (1982), deleting a small number of variables from the analysis should not, in most cases, greatly alter the clusters if these clusters are real and not mere artifacts of the technique being used. Furthermore, clustering allows researchers to process large data sets efficiently and can accommodate both quantitative and qualitative variables if conducted on factorial coordinates. It can also accommodate small sample sizes without restricting the number of covariates in the model. Moreover, using well-consolidated techniques of cluster profiling (e.g., Cohen et al. 1977; Jobson, 1992), a direct description of the clusters can be obtained, making it simpler to interpret and use the results compared to other methods.

Limitations

We are aware that the subgroup analysis strategy here proposed might have some limitations. First, the algorithm chosen for a given cluster analysis can have important implications for analytical results (e.g., Aldenderfer & Blashfield, 1984). Almost all clustering algorithms will produce partitions for any dataset, regardless of whether the resulting clustering actually represents a significant cluster structure or not (Milligan, 1981). Since each algorithm is associated with potential shortcomings, no one particular algorithm can be considered the best. However, several studies have demonstrated that some algorithms are more appropriate than others for particular kinds of data (e.g., Jobson, 1992; Lebart, Morineau, & Tabard, 1977; Lebart, Morineau, & Warwick, 1984; Milligan, 1980). Because Lebart, Morineau, and Tabard (1977) demonstrated that Ward's algorithm is compatible with MCA, as both are based on a similar optimization criterion (inertia definition), we also used Ward's algorithm for this study. However, other less appropriate algorithms may have given different results. Rather than focusing on a particular algorithm to perform our cluster analysis, future studies might follow a new line of research that employs consensus clustering to obtain a single consensus solution (e.g.,

Goder & Filkov, 2008; Topchy et al. 2004; Strehl & Gosh, 2002). This methodology might avoid the potential shortcomings of various clustering methods and improve the reliability of our clusters solutions.

Single Cluster Solution vs. Averaged Cluster Solutions

Selecting an optimal cluster solution also involves some subjectivity on the part of the researcher. Although several reasonable solutions can be identified, the optimal number of subgroups often depends on the specific technique used to determine the number of clusters used for the statistical adjustment (Chiu, Fang, Chen, Wang, & Yeris, 2001; Fraley & Raftery, 1998; Salvador & Chan, 2004; Tibshirani, Walther, & Hastie, 2001; D'Attoma & Liberati, 2011). To the best of authors' knowledge, no optimal criterion takes into account balance and common support. Therefore, we estimated the average treatment effect for several k-cluster solutions, rather than choosing a particular optimal solution, in addition to providing the results from a single, hand-picked cluster solution. That was possible because, in an evaluation context, the number of all possible k-clusters solutions is limited since partitions without sufficient common support and/or covariate balance are excluded. Therefore, we were able to avoid the potential shortcomings in using various criteria to select an optimal number of groups by providing results from the weighted ATE. This is a good alternative to using optimization criteria because it simultaneously allows for maintaining criteria for selecting appropriate cluster solutions and some flexibility in interpreting the results to account for statistical error.

Theoretically, several cluster-based solutions may be equally effective in reducing bias. However, when applying this method in a program evaluation, it is important to obtain a final cluster solution that balances covariates without discarding too many units. The measure of GI permits researchers to gauge the trade-off between covariate balance, which improves with more clusters, and common support, which improves with fewer clusters, when selecting the optimal cluster solution from the dendrogram.

Conclusions

Although the results from this study do not permit us to definitively claim that the cluster-based method was *significantly more* effective in reducing bias than propensity score subclassification or IPTW, we were able to show strong evidence that it reduces bias and retains participants as well as two of the better propensity score adjustment methods that many researchers use. Furthermore, the cluster-based approach may reduce some of the limitations that propensity score adjustment methods face. For those wanting to use the cluster-based approach to reduce selection bias in nonrandomized studies, this study offers two possible strategies: 1) compute the ATE of all possible k-cluster solutions that permit sufficient common support and global balance; and 2) select a single cluster solution that might produce the most bias reduction. The first strategy minimizes the researcher's choices in conducting the analysis, while still providing an acceptable reduction in bias reduction. While the second strategy provides a greater reduction in bias with a little more effort.

References

Aldenderfer, M. S., & Blashfield, R.K., (1984). *Cluster analysis*. Beverly Hills, CA: Sage Publications.

Anderberg, M.R. (1973). *Cluster Analysis for Applications*. Academic Press, Inc., London.

Austin, P.C. (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine, 33*, 1057-1069. doi: 10.1002/sim.2328

Austin, P.C. (2011). An Introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(1), 399-424.

Austin, P.C., & Mamdani, M.M. (2006). A comparison of propensity score methods: A case study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine, 25*, 2084-2106. doi: 10.1002/sim.2328

Austin, P.C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine, 34*, 3661-3679. doi: 10.1002/sim.6607

Bai, H. (2013). A bootstrap procedure of propensity score estimation. *The Journal of Experimental Education,81,* 157-177. doi: 101080/00220973.2012.700497

Benzecri, J.P. (1973). *L'analyse des données*. Paris: Dunod.

Camillo, F., D'Attoma, I. (2012). %GI : A SAS Macro for Measuring and Testing Global Imbalance of Covariates within Subgroups. Journal of Statistical Software, Volume 51, Code Snippet 1, pp. 1-19.

Chiu, T., Fang, D., Chen, J., Wang, Y. & Jeris, C., (2001). A robust and scalable, clustering algorithm for mixed type attribute in large database environment. In: *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 263-68). San Francisco, CA: ACM.

Christakis, N.A., & Iwashyha, T.I., (2003). The health impact of health care on families: A matched cohort study of hospice use by decedents and mortality outcomes in surviving, widowed spouses. *Social Science and Medicine, 57,* 465-475.

Clark, M.H. (2011, November). *A Comparison of Bias Reduction Methods on Educational Outcomes*. Paper presented at the American Evaluation Association Convention, Anaheim, CA.

Cochran, W.G., (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, *24*, 205-213.

Cohen, A., Gnanadesikan, R., Kettenring, J.R. and Landwehr, J.M. (1977). Methodological developments in some applications of clustering. In *Applications of statistics* (P.R. Krishnaiah, ed.). Amsterdam: North Holland Publishing Co.

D'Attoma, I., & Camillo, F. (2011). A multivariate strategy to measure and test Global Imbalance in observational studies. *Expert Systems with Applications,38,* 3451-3460.

D'Attoma I., & Liberati, C. (2011). An optimal cluster-based approach for subgroup analysis using information complexity criterion. *International Journal of Business Intelligence and Data Mining, 6* (4), 402-425.

Dehejia, R.H. & Wahba, S., (1999). Causal effects in nonexperimental studies: Re-evaluating the evaluation of training programs. *Journal of the American Statistical Association, 94,* 1053-1062.

Diamond, A., & Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observation studies. *Review of Economics and Statistics.* 95(3), 932-945.

Duda, R.O., Hart, P.E. and Stork, D.G. (2001). *Pattern classification.* Hoboken, NJ: John Wiley & Sons.

Estadella, J.D., Aluja-Banet, T. & Thio-Henestrosa, S. (2005). Distribution of the inter and intra inertia in conditional MCA. *Computational Statistics*, 20: 449-463.

Everitt, B.S. (1993). *Cluster analysis*. London, UK: Arnold.

Everitt, B.S., Landan, S., Leese, M. and Stahl, D. (2011). Cluster Analysis. Wiley Online Library, 5th edition.

Fraley, C. & Raftery, A., (1998). How many clusters? Which clustering method? Answer via model-based cluster analysis. *Computer Journal, 41(8),* 578-588.

Freedman, D.A., & Berk, R.A. (2008). Weighting regressions by propensity scores. *Evaluation Review, 32*, 392-409. doi: 10.1177/0193841X08317586

Gerfin, M., & Lechner, M., (2002). A microeconometric evaluation of the active labour market policy in Switzerland. *The Economic Journal*, *112*, 854-893.

Garrido, M.M., Kelley, A.S., Paris, J., Rosa, K., Meier, D.E., Morrison, R.S., Aldridge, M.D. (2014). Methods for constructing and assessing propensity scores. *Health Services Research, 49*, 1701-1720.

Gibson, C.M., (2003). Privileging the participant: The importance of subgroup analysis in social welfare evaluations. *American Journal of Evaluation, 24 (4),* 443-469.

Goder, A., & Filkov, V. (2008). Consensus clustering algorithms: Comparison and refinement. *Proceedings of the Ninth Workshop on Algorithm Engineering and Experiments*. San Francisco, CA: Society for Industrial and Applied Mathematics.

Green, P.E., & Krieger, A.M., (1995). Alternative approaches to cluster-based market

segmentation. *Journal of Market Research Society, 37,* 221-229.

Greenacre, M.J. (1984). *Theory and applications of correspondence analysis*. London, UK: Academic Press.

Hansen, B.B., & Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*, 23(2), 219-236.

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. doi: 10.1037/a0019623

Harknett, K., (2006). Does receiving an earnings supplement affect union formation? Estimating effects for program participants using propensity score matching. *Evaluation Review*, *30*,741-778.

Hastie, T., Tibishirani, R. and Friedman, J. (2001). *The elements of statistical learning: Data mining, inference and prediction.* New York, NY: Springer.

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(3)*, 199-236.

Hong, G., & Raudenbush, S.W., (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational and Policy Analysis*, *27(3),* 205-224.

Hong, G. (2007). Marginal mean weighting adjustment for selection bias. Toronto, Canada: Ontario Institute for Studies in Education of the University of Toronto. Unpublished manuscript.

Horvitz, D.G. & Thompson, D.J. (1952). "A generalization of sampling without replacement from a finite universe." *Journal of the American Statistical Association*, 47, 663-685.

Iacus, S.M., King, G., & Porro, G. (2011). Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association, 106* (493), 345-361.

Imbens, G.W., (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics, 86*, 4-29.

Jobson, J.D. (1992). *Applied Multivariate Data Analysis. Volume II: Categorical and Multivariate Methods*. New York, NY: Springer.

Jolliffe, I.T., Jones, B., & Morgan, B.J.T. (1982). Utilising clusters: A case study involving the elderly. *Journal of the Royal Statistical Society, A*, 145, 224-236.

Kang, J.D.Y., & 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.

King, G. & Nielsen, R. (2016). Why propensity scores should not be used for matching. Working paper retrieved from j.mp/PScore.

Lanehart, R.E., De Gil, P.R., Kim, E.S., Bellara, A.P., Kromrey, J.D. and Lee, R.S. (2012). Propensity Score Analysis and Assessment of Propensity Score Approaches Using SAS Procedures. SAS Global Forum Paper 314-2012.

Lebart, L., Morineau, A., & Tabard, N. (1977). *Technique de la Description Statistique*. Paris: Dunod.

Lebart, L., Morineau, A., & Warwick, K.M. (1984). *Multivariate descriptive statistical analysis: Correspondence analysis and related techniques for large matrices*. New York, NY: John Wiley & Sons.

Lebart, L., Morineau, A., & Piron, M. (1997). *Statistique exploratoire multidimensionelle*. Paris: Dunod.

Li, Q., Maasoumi, E., & Racine, J.S. (2009). A nonparametric test for equality of distributions with mixed categorical and continuous data. *Journal of Econometrics, 148*, 186-200.

Lunceford, J.K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimations of causal treatment effects: A comparative study. *Statistics in Medicine, 23*, 2937-2960.

Milligan, G.W. (1980). An examination of the effect of six types of error perturbation of fifteen clustering algorithms. *Psychometrika, 45(3),* 325-42.

Milligan, G.W., (1981). A Monte Carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika, 46(2)*, 187-99.

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

Peck, L.R., (2005). Using cluster analysis in program evaluation. *Evaluation Review, 29(2),* 178-196.

Peck, L.R., (2007). What are the effects of welfare sanction policies? Or, using propensity

scores as a subgroup indicator to learn more from social experiments. *American Journal of Evaluation, 28,* 256-274.

Peck, L.R., (2011, November). *A new strategy for reducing selection bias in non-experimental evaluations.* Paper presented at the American Evaluation Association Convention, Anaheim, CA.

Peck, L.R., Camillo, F., & D'Attoma, I. (2010). A promising new approach to eliminating selection bias. *The Canadian Journal of Program Evaluation, 24*(2), 31-56.

Peck, L.R., D'Attoma, I., Camillo, F. , Guo, C. (2012). A new strategy for reducing selection bias in non-experimental evaluations, and the case of how public assistance receipt affects charitable giving. *Policy Studies Journal, 40*, 601-625.

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

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

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

Rubin, D.B., (1973). Matching to remove bias in observational studies. *Biometrics, 29*, 159-183.

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

Salvador, S. & Chan, P. (2004). Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *Proceedings of the 16th IEEE International Conference on Tools with AI (ICTAI)*. Los Angeles, CA: IEEE Computer Society.

Shadish, W.R., Clark, M.H., & Steiner, P.M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignment. *Journal of the American Statistical Association, 103*, 1334-1356.

Shadish, W.R. & Steiner, P.M. (2010). A primer on propensity score analysis. Newborn and Infant Nursing Reviews, 10(1), 19-26.

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

Smith, J., & Todd, P., (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics, 125*, 305-353.

Stone, C. A., & Tang, Y. (2013). Comparing propensity score methods in balancing covariates and recovering impact in small sample educational program evaluations. *Practical Assessment, Research & Evaluation, 18*(13), 1-12.

Strehl, A., & Ghosh, J. (2002). Cluster ensembles: A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3, 583-617.

Tebes, J.K., Feinn, R., Vanderploeg, J.J., Chinman, M.J., Shepard, J., Brabham, T.,…,Connel, C., (2007). Impact of a positive youth development program in urban after-school settings on the prevention of adolescent substance use. *Journal of Adolescent health, 41(3),* 239-247.

Thoemmes, F. J., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research, 46*, 90-118.

Tibshirani, R., Walther, G., & Hastie, T., (2001). Estimating the number of clusters in a data set via the Gap Statistic. *Journal of the Royal Statistical Society (Series B), 63(2),* 411-423.

Topchy, A., Jain, A. K., & Punch, W. (2003). Clustering ensembles: Models of consensus and weak partitions. In *IEEE International Conference on Data Mining, ICDM 03 & SIAM International Conference on Data Mining*. Los Angeles, CA: IEEE Computer Society.

Yoshikawa, H., Rosman, E.A., & Hsueh, J., (2001). Variation in teenage mothers', experiences of child care and other components of welfare reform: Selection processes and developmental consequences. *Child Development, 72*, 299-317.

Zanutto, E.L., (2006). A comparison of propensity score and linear regression analysis of complex survey data. *Journal of Data Science, 4*, 67-91.

Table 1. Means, Standard Deviations and Correlations for Covariates and Outcome Variables

	Nonrandomized Samples				Randomized Samples			
Variable	**Mean (Standard Deviation)**	**Correlation with Vocabulary Training**	**Correlation with Vocabulary Posttest**	**Correlation with Mathematics Posttest**	**Mean (Standard Deviation)**	**Correlation with Vocabulary Training**	**Correlation with Vocabulary Posttest**	**Correlation with Mathematics Posttest**
Vocabulary Pretest	21.01 (5.20)	.17	.47	.11	20.69 (4.64)	-.05	.26	.24
Mathematics Pretest	6.29 (2.73)	-.09	.15	.45	6.21 (2.79)	-.02	.17	.27
Number of Previous Mathematics Courses	2.09 (1.12)	-.02*	-.06*	.33*	1.86 (0.92)	-.07*	-.04*	.36*
Like Mathematics	4.86 (2.42)	-.36	-.29	.47	4.33 (2.28)	.02*	-.04*	.30*
Like Literature	6.40 (2.22)	.16	.23	-.23	6.34 (2.13)	-.04	.09	-.17
Prefer Literature over Mathematics	2.30 (.898)	.38*	.42*	-.43*	2.39 (0.86)	-.04*	.02*	-.27*
Extraversion	32.32 (7.82)	.09	.01	-.16	32.19 (7.57)	.04	0.00	-.21
Agreeableness	40.50 (4.85)	.10	.12	-.08	39.54 (5.24)	-.11	-.06	-.03
Conscientiousness	34.83 (5.96)	-.13	-.19	-.04	34.92 (5.37)	.05	.09	-.02

Emotional Stability	29.81 (7.74)	-.02	-.10	-.12	28.98 (8.14)	.14	.13	-.05
Openness to New Experiences	36.32 (5.08)	.05	.20	.05	35.65 (5.14)	-.09	.07	.07
Mathematics Anxiety	58.03 (22.21)	.00	-.05	-.14	60.09 (21.25)	-.05	-.05	-.16
Depression	4.83 (6.18)	.10*	.05*	.16*	4.84 (4.77)	-.09*	-.05*	.09*
Caucasian	0.56 (0.50)	.18**	.32	-.07	0.58 (0.49)	.02**	.15	.01
African American	0.39 (0.49)	-.14**	-.30	-.02	0.34 (0.48)	-.02**	-.19	-.04
Age	21.67 (5.43)	.10*	.08*	-.22*	20.85 (4.69)	.06	.11	-.19
Male	0.24 (0.43)	-.07**	.06	.14	0.28 (0.45)	.05**	.08	.08
Married	0.08 (0.27)	.00**	-.07	-.16	0.03 (0.17)	.03**	.14	-.05
Mother's Education (years)	14.26 (2.30)	.01	.09	-.02	14.09 (2.39)	-.05*	-.04*	.10*
Father's Education (years)	14.50 (2.74)	.01	.11	.07	14.19 (2.50)	-.01	.04	.16
Hours of College Credit	31.36 (33.48)	.03	.10	.10	26.69 (30.15)	-.13*	*-.07	.23*
STEM Major	0.20 (0.40)	-.19**	-.17	.30	0.27 (0.44)	.04**	.03	.11
ACT Score	21.66 (4.75)	.03	.34	.42	21.05 (4.06)	-.05	.24	.42

High School GPA	3.10 (0.56)	-.04	.00	.40	2.99 (0.63)	-.05	.06	.33
College GPA	2.77 (0.69)	-.03	.06	.22	2.60 (0.73)	-.02	.13	.28
Vocabulary Posttest	13.36 (5.64)	.78		-.27	12.08 (5.29)	.77		-.27
Mathematics Posttest	9.26 (4.57)	-.53	-.27		9.29 (4.01)	-.52	-.27	

Note: Point biserial correlations were used to compare continuous covariates with vocabulary training condition and dichotomous covariates with outcome variables. Pearson correlations were used to compare continuous covariates with outcome variables. *Spearman's rho was used when variables were ordinal or not normally distributed. **Phi coefficients were used when both variables were dichotomous.

Table 2. Mathematics outcome

	Treatment Effect (standard error)	**Absolute bias**
Covariate-adjusted randomized experiment	3.98 (.34)	.00
Unadjusted randomized experiment	4.19 (.44)	.21
Unadjusted quasi-experiment	5.01 (.55)	1.03

Note. The treatment effect is the difference between the average number of items correct from the treatment and control groups. The absolute bias is the difference between each treatment effect and the treatment effect from the Covariate-adjusted randomized experiment.

Table 3. Propensity score adjustment results for the mathematics outcome.

PS Adjustment	**Treatment Effect (standard error)*** **	**Absolute bias**	**Percent Bias Reduction**
PS stratification	3.72(.57)	.26	75%

PS stratification plus covariates	4.05(.59)	.07	93%
IPTW	3.60(.54)	.38	63%
IPTW plus covariates	3.72(.36)	.26	75%

Note. The treatment effect is the difference between the average number of items correct from the treatment and control groups. The absolute bias is the difference between each treatment effect and the treatment effect from the Covariate-adjusted randomized experiment.

Table 4. Global balance of all covariates.

Mathematics training	**Vocabulary training**	**Global Imbalance**	**Interval**	**Balance**	**N**
N=79	N=131	0.03646	(0;0.02558)	NO	210

Table 5. Balance summary within the selected k-cluster solutions.

k-cluster solution	Balance		
	Alpha=0.05	**Alpha=0.1**	**Alpha=0.01**
k=2	1 unbalanced group	1 unbalanced group	1 unbalanced group
k=3	1 unbalanced group	1 unbalanced group	All balanced
k=4	1 unbalanced group	1 unbalanced group	1 unbalanced group
k=5	1 unbalanced group	1 unbalanced group	All balanced
k=6	1 unbalanced group	1 unbalanced group	1 unbalanced group
k=7	1 unbalanced group	2 unbalanced groups	1 unbalanced group
k=8	1 unbalanced group	2 unbalanced groups	1 unbalanced group
k=9	All balanced	All balanced	All balanced
k=10	All balanced	All balanced	All balanced
k=11	All balanced	All balanced	All balanced
k=12	All balanced	All balanced	All balanced
k=13	All balanced	All balanced	All balanced
k=14	All balanced	All balanced	All balanced
k=15	All balanced	All balanced	All balanced
k=16	All balanced	All balanced	All balanced
k=17	No common support	No common support	No common support
k=18	No common support	No common support	No common support
k=19	No common support	No common support	No common support
k=20	No common support	No common support	No common support

Table 6. ATEd and percent bias reduction by k-cluster partition for the mathematics outcome.

k-cluster partition	**Overall Math ATEd (standard error)**	**Absolute Bias**	**Percent Bias Reduction**
9	4.4 (.61)	0.42	59%
10	4.36 (.64)	0.38	63%
11	4.49(.67)	0.51	50%
12	4.08(.56)	0.10	90%
13	4.02(.55)	0.04	96%
14	4.05(.56)	0.07	93%
15	3.95(.45)	0.03	97%
16	3.94(.45)	0.04	96%

Table 7. ATEs and balance in the 13-cluster solution for the mathematics outcome.

Cluster	N	Mathematics training	Vocabulary training	GI	95%CI	Balance	Treatment effect (std)	MIC
1	14	6	8	0.16	(0,0.33)	Yes	6.29(2.03)	0.0640
2	14	9	5	0.26	(0,0.33)	Yes	2.68(2.41)	0.1058
3	17	11	6	0.21	(0,0.28)	Yes	5.51(1.61)	0.0796
4	14	2	12	0.22	(0,0.34)	Yes	4.41(3.37)	0.0834
5	18	4	14	0.21	(0,0.25)	Yes	5.92(2.1)	0.0861
6	23	10	13	0.10	(0,0.21)	Yes	4.55(1.06)	0.0370
7	11	6	5	0.29	(0,0.43)	Yes	1.7(2.08)	0.1126
8	29	1	28	0.14	(0,0.16)	Yes	1.17(3.31)	0.0523
9	8	6	2	0.34	(0,0.50)	Yes	9(2.75)	0.1696
10	15	7	8	0.15	(0,0.30)	Yes	0.83(2.05)	0.0605
11	10	3	7	0.22	(0,0.36)	Yes	5.19(2.77)	0.1293
12	18	6	12	0.17	(0,0.25)	Yes	4.41(1.85)	0.0719
13	19	8	11	0.12	(0,0.24)	Yes	4.43(1.83)	0.0505

Table 8. ATEs and balance in the 15-cluster solution for the mathematics outcome.

Cluster	N	Mathematics training	Vocabulary training	GI	95%CI	Balance	Treatment effect (std)	MIC
1	14	5	9	0.19	(0,0.33)	Yes	6.29(2.12)	0.0762
2	14	9	5	0.26	(0,0.33)	Yes	2.68(2.42)	0.1058
3	17	11	6	0.21	(0,0.28)	Yes	5.51(1.61)	0.0796
4	14	1	13	0.22	(0,0.34)	Yes	1.00(4.79)	0.0811
5	10	2	8	0.34	(0,0.36)	Yes	10.12(1.87)	0.1991
6	10	4	6	0.33	(0,0.44)	Yes	2.67(2.95)	0.1417
7	23	10	13	0.10	(0,0.21)	Yes	4.55(1.06)	0.0370
8	10	6	4	0.28	(0,0.41)	Yes	3.25(1.74)	0.1227
9	27	1	26	0.14	(0,0.17)	Yes	1.03(3.39)	0.0539
10	8	6	2	0.33	(0,0.50)	Yes	9.00(2.75)	0.1696
11	8	5	3	0.19	(0,0.45)	Yes	-0.93(2.66)	0.1137
12	8	2	6	0.17	(0,0.46)	Yes	3.67(4.13)	0.0973
13	10	3	7	0.22	(0,0.36)	Yes	5.19(2.77)	0.1293
14	18	6	12	0.17	(0,0.25)	Yes	4.41(1.86)	0.0719
15	19	8	11	0.13	(0,0.25)	Yes	4.07(1.65)	0.0493

Table 9. ATEs and balance in the 16-cluster solution for the mathematics outcome.

Cluster	N	Mathematics training	Vocabulary training	GI	95%CI	Balance	Treatment effect (std)	MIC
1	14	5	9	0.19	(0,0.33)	Yes	6.29(2.12)	0.0762
2	14	9	5	0.26	(0,0.33)	Yes	2.69(2.42)	0.1058
3	7	4	3	0.31	(0,0.50)	Yes	6.75(2.66)	0.1983
4	10	7	3	0.36	(0,0.41)	Yes	4.52(2.26)	0.1682
5	14	1	13	0.22	(0,0.34)	Yes	1.00(4.79)	0.0811
6	10	2	8	0.34	(0,0.36)	Yes	10.12(1.87)	0.1991
7	10	4	6	0.33	(0,0.44)	Yes	2.67(2.95)	0.1471
8	23	10	13	0.10	(0,0.21)	Yes	4.55(1.00)	0.0370
9	10	6	4	0.28	(0,0.44)	Yes	3.25(1.74)	0.1227
10	27	1	26	0.14	(0,0.17)	Yes	1.04(3.40)	0.0539
11	8	6	2	0.34	(0,0.50)	Yes	9.00(2.75)	0.1696
12	8	5	3	0.19	(0,0.45)	Yes	-0.93(2.66)	0.1137
13	8	2	6	0.17	(0,0.46)	Yes	3.67(4.13)	0.0973
14	10	3	7	0.22	(0,0.36)	Yes	5.20(2.77)	0.1293
15	18	6	12	0.17	(0,0.25)	Yes	4.42(1.86)	0.0719
16	19	8	11	0.13	(0,0.25)	Yes	4.07(1.65)	0.0493

Table 10. Vocabulary outcome

	Treatment effect (standard error)* **	**Absolute bias**
Covariate-adjusted randomized experiment	8.25 (.37)	.00
Unadjusted randomized experiment	8.11 (.44)	.14
Unadjusted quasi-experiment	9.00 (.51)	.75

*Note. *** The treatment effect is the difference between the average number of items correct from the treatment and control groups. The absolute bias is the difference between each treatment effect and the treatment effect from the Covariate-adjusted randomized experiment.*

Table 11. Propensity score adjustment results for the vocabulary outcome.

PS Adjustment	Treatment Effect (standard error)***	Absolute bias	Percent Bias Reduction
PS stratification	8.15(.60)	.10	87%
PS stratification plus covariates with strata	8.32(.49)	.07	91%
IPTW	8.22(.49)	.03	96%
IPTW plus covariates	8.18(.38)	.07	91%

*Note. *** The treatment effect is the difference between the average number of items correct from the treatment and control groups. The absolute bias is the difference between each treatment effect and the treatment effect from the Covariate-adjusted randomized experiment.*

Table 12. ATEd and percent bias reduction by k-cluster partition for the vocabulary outcome.

k-cluster partition	Overall Vocab ATEd (standard error)	Absolute Bias	Percent Bias Reduction
9	8.04 (.43)	.20	72%
10	8.10 (.44)	.15	80%
11	7.95 (.51)	.30	60%
12	8.00 (.50)	.25	67%
13	8.04 (.51)	.21	72%
14	8.01 (.50)	.24	68%
15	8.15 (.42)	.10	87%
16	8.27 (.42)	.02	97%

Table 13. ATEs and balance in the 16-cluster solution for the vocabulary outcome.

Cluster	N	Mathematics training	Vocabulary training	GI	95%CI	Balance	Treatment Effect (std)	MIC
1	14	5	9	.19	(0,0.33)	Yes	9(1.89)	0.0762
2	14	9	5	.26	(0,0.33)	Yes	11.91(1.61)	0.1058
3	7	4	3	.31	(0,0.50)	Yes	10.42(0.93)	0.1983
4	10	7	3	.36	(0,0.41)	Yes	8.86(1.77)	0.1682
5	14	1	13	.22	(0,0.34)	Yes	10(4.42)	0.0811
6	10	2	8	.34	(0,0.36)	Yes	9.75(2.53)	0.1991
7	10	4	6	.33	(0,0.44)	Yes	7.25(2.62)	0.1471
8	23	10	13	.1	(0,0.21)	Yes	6.87(1.38)	0.0370
9	10	6	4	.28	(0,0.44)	Yes	5.67(2.06)	0.1227
10	27	1	26	.14	(0,0.17)	Yes	7.31(4.13)	0.0539
11	8	6	2	.34	(0,0.50)	Yes	1.33(2.89)	0.1696
12	8	5	3	.19	(0,0.45)	Yes	9.8(1.07)	0.1137
13	8	2	6	.17	(0,0.46)	Yes	10.17(2.33)	0.0973
14	10	3	7	.22	(0,0.36)	Yes	5.71(1.85)	0.1293
15	18	6	12	.17	(0,0.25)	Yes	9.08(1.92)	0.0719
16	19	8	11	.13	(0, 0.25)	Yes	8.91(1.40)	0.0493

Figure 1. The dendrogram from the cluster-based method results.

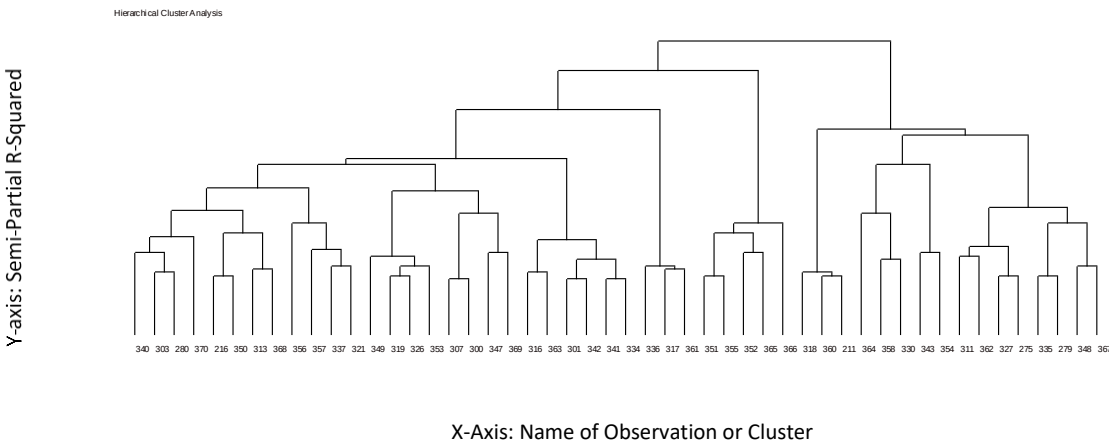
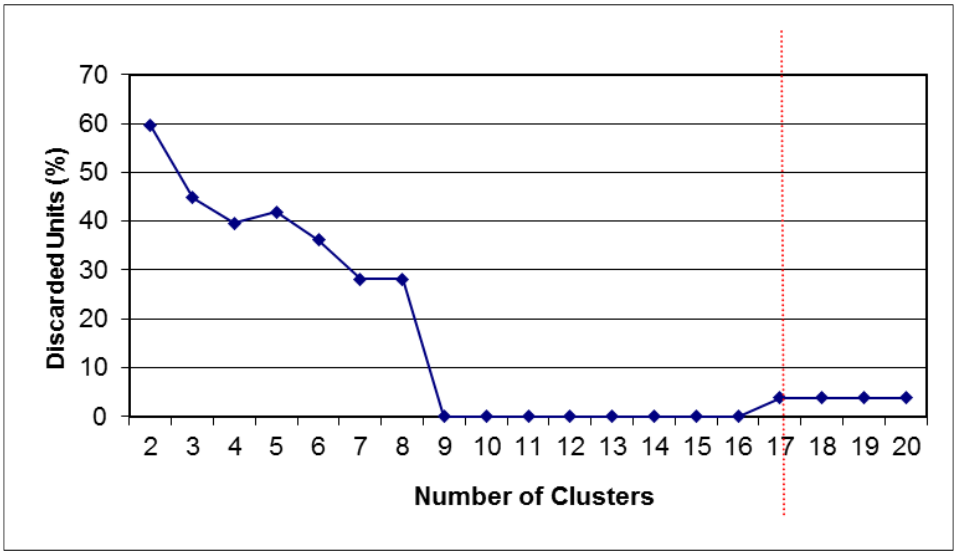


Figure 2. The percent of units discarded for each k-cluster solution.



Note. Cluster solutions to right of the red vertical line did not satisfy the common support criterion.

Figure 3. Percent bias reduction in the selected partitions for the mathematics outcome.

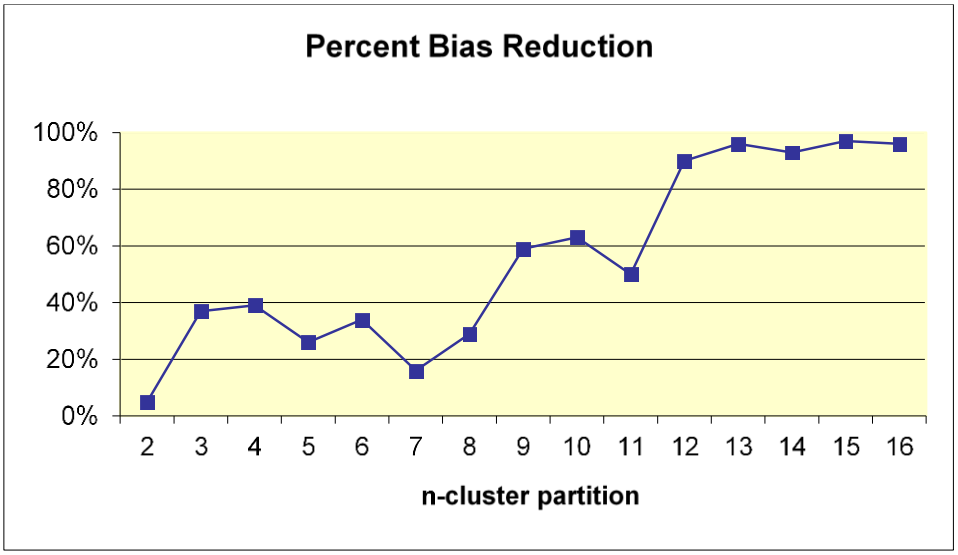
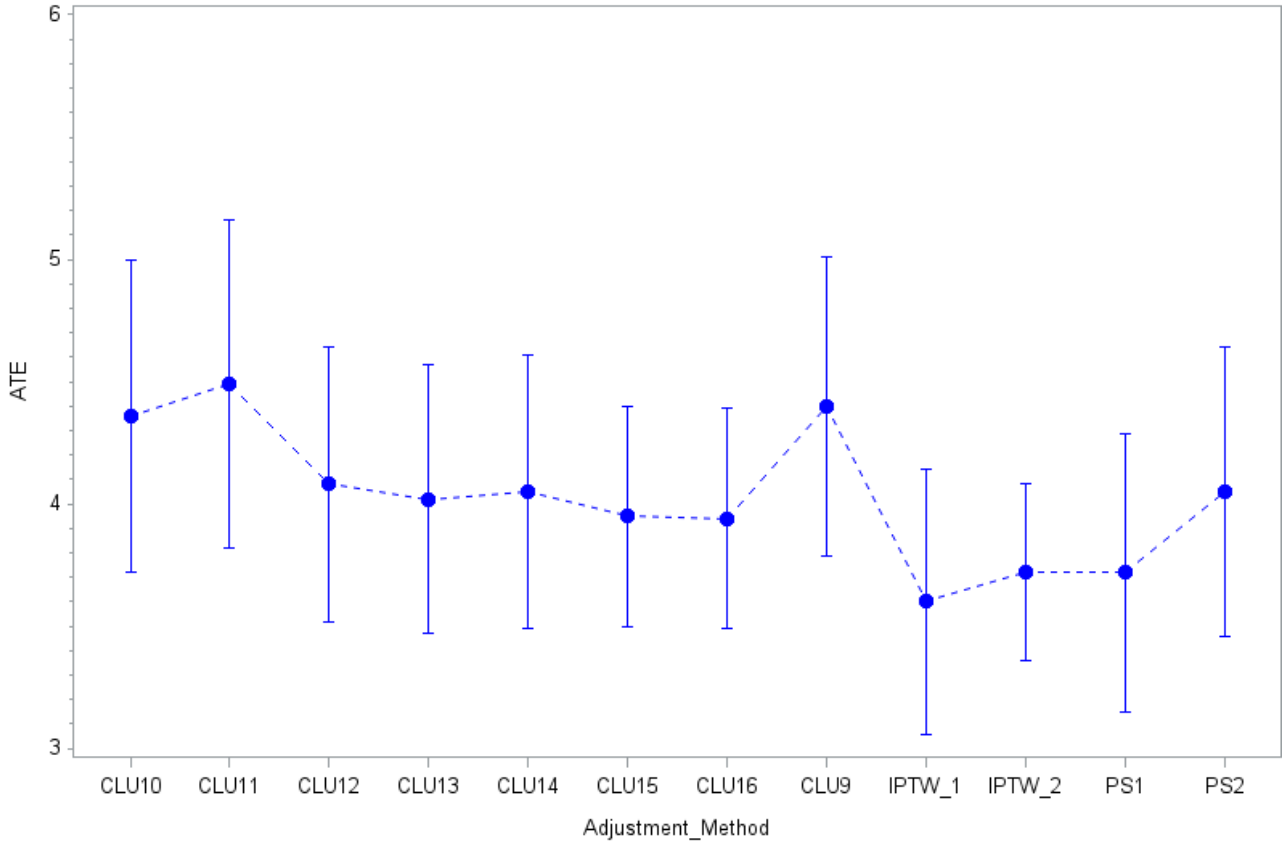
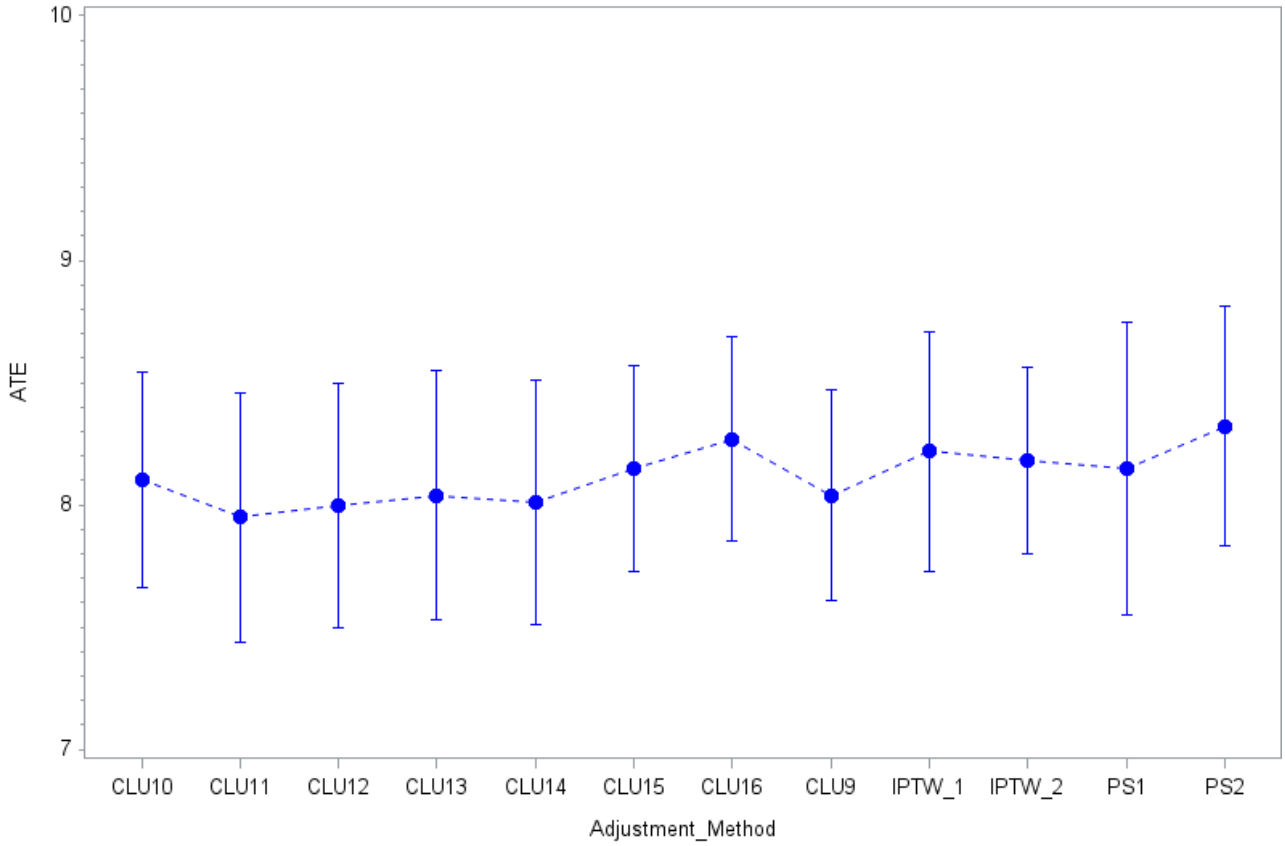


Figure 4. ATE_d with standard error bars for the mathematics outcome.



PS1=PS subclassification
PS2=PS subclassification plus covariates with strata
IPTW_1=inverse probability of treatment weighting
IPTW_2=inverse probability of treatment weighting plus covariates

Figure 5. ATE_d with standard error bars for the vocabulary outcome.



PS1=PS subclassification
PS2=PS subclassification plus covariates with strata
IPTW_1=inverse probability of treatment weighting
IPTW_2=inverse probability of treatment weighting plus covariates

Appendix A

Standardized bias estimates before and after propensity score adjustments and the percent of selection bias reduced by each method.

Covariate	$SB_{unadjusted}$	$SB_{subclass}$	$\%\ BR_{subclass}$	SB_{IPTW}
Vocabulary pretest	-0.353	-0.046	87.055	-0.014
Mathematics pretest	0.186	0.016	91.544	0.014
Previous mathematics courses	0.272	-0.026	90.595	-0.006
Liked mathematics	0.782	-0.006	99.207	0.001
Preferred literature over mathematics	-0.855	0.039	95.430	-0.003
Extraversion	-0.191	-0.027	86.080	-0.038
Conscientiousness	0.262	-0.038	85.323	-0.002
Emotional Stability	0.031	-0.038	-20.422	-0.050
Mathematics anxiety	-0.006	-0.024	-326.905	-0.024
Depression	0.029	-0.031	-7.085	-0.014
Caucasian	-0.410	-0.068	83.468	-0.006
Male	0.169	0.010	94.118	-0.038
Mother's Education	-0.205	-0.115	43.818	-0.089
Mother's Education^2	-0.200	-0.107	46.455	-0.084
Father's Education	-0.018	-0.094	-417.044	-0.043
Math-intensive major	0.521	0.061	88.254	-0.023
College GPA	0.054	0.004	93.303	0.014
College GPA^2	0.100	0.008	92.276	0.016

Notes: $SB_{unadjusted}$ is the standardized bias of the covariates before propensity score methods were applied, $SB_{subclass}$ is the bias after propensity score subclassification, $\%\ BR_{subclass}$ is the percent of bias reduced after subclassification, $SB_{subclass + cov}$ is the bias after subclassification and accounting for the covariates, $\%\ BR_{subclass + cov}$ is the percent of bias reduced after subclassification and accounting for the covariates.

Covariate bias was estimated using a formula similar to that used by Rosenbaum and Rubin (1985). Because the sample sizes were not equal (n_m = 79 and n_v = 131), the standardized measure of bias (*SB*) used the pooled standard deviation, rather than the average standard deviation of the groups: $SB = \left(M_m - M_v\right) / \sqrt{\left(s_m^2 (n_m - 1) + s_V^2 (n_V - 1)\right) / \left(n_m + n_v - 2\right)}$, where M_m is the mean of the mathematics training group, M_v is the mean of the vocabulary training group, s_m^2 is the variance of the mathematics training group, s_v^2 is the variance of the vocabulary training group, and n_m and n_v are the sample sizes of the mathematics and vocabulary training groups, respectively. Odds ratios computed for the categorical covariates were converted to *SB* using a formula similar to Hasselblad and Hedges's (1995): $SB = (\ln(OR)\sqrt{3}) / \pi$.