Contact email: aburnett@mathematica-mpr.com

Title: The Performance of Balance Diagnostics for Propensity Score Matched Samples in Multilevel Settings

First choice of conference section: Research Methods

Authors: Alyson Burnett, Mathematica, aburnett@mathematica-mpr.com

Laura Stapleton, University of Maryland, lstaplet@umd.edu

Presenting author: Alyson Burnett

Background:

Performing diagnostics is an essential step required for studies using propensity score (PS) matching. Ho et al. (2007) describe the diagnostic step as an iterative process that involves testing many PS models and matching methods, modifying them as necessary, and selecting an approach that will result in unbiased treatment effect (TE) estimates. PS diagnostics evaluate whether units in the treatment and control groups have similar means and distributions of the measured covariates, a feature known as balance. Demonstrating that a study's samples are well balanced is required for any quasi-experiments wishing to meet the What Works Clearinghouse's standards, a designation that has implications for education programs seeking federal grant funding.

Although researchers have recently extended PS matching to multilevel settings, it is not yet clear how to apply diagnostic procedures to nested data. Methodological researchers have investigated the modeling and conditioning approaches used in multilevel PS methods but have not yet investigated diagnostic approaches (e.g., Arpino & Cannas, 2016; Arpino & Mealli, 2011; Rickles & Seltzer, 2014; Thoemmes & West, 2011). This simulation study sought to expand the literature on methods for multilevel PS matching and provide guidance to researchers wanting to assess balance in multilevel settings.

Purpose:

This study tested several possible measures for evaluating balance of multilevel PS-matched samples. The two primary measures for evaluating balance of covariates in single-level studies include the absolute standardized bias (ASB), or mean difference, between treatment and control groups, and the ratio of treatment and control variances (Thoemmes & Kim, 2011; Austin, 2009; Rubin, 2001).

In a multilevel PS matching study, ASB and variance ratios may be pooled across clusters, ignoring cluster membership ("pooled balance"), or they may be calculated separately for each cluster and then summarized ("within-cluster balance"). Given the lack of investigation of balance assessment in multilevel PS applications, this study sought to answer the following questions: Which pooled and within-cluster measures of variance ratios and ASB are most related to bias in the treatment effect estimate? Does this vary according to intracluster correlation coefficients (ICCs) of the unit-level covariates, cluster size, PS model, or matching method?

Based on prior research, we hypothesized that the preferred balance measure should depend on several factors, including the size of the clusters, the value of the ICCs of the unit-level covariates, the extent of the misspecification of the PS model, and the matching method. These factors were manipulated in the Monte Carlo simulation to better understand the conditions in which different balance measures would be useful.

Design:

Using a Monte Carlo simulation, the study compared the correlations between various balance measures with bias in the treatment effect estimate in several multilevel contexts. The simulation required two data generation models—the PS model, which generated the probability of being treated using logistic regression, and the TE model, which generated the value on the outcome variable using linear regression. The parameters in both models were based on a separate empirical analysis that examined the effect of kindergarten retention on

reading outcomes using the ECLS-K: 2011 data (NCES, Tourangeau et al., 2015). The propensity scores were generated with the following model:

$$
\begin{aligned}
&\text{logit}(\mathrm{T_{ij}}) = \beta_{0j} + \beta_{Rj} X_{Rij} + \beta_{Mj} X_{Mij} + \beta_{Aj} X_{Aij} \\
&\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j} \\
&\beta_{Rj} = \gamma_{R0} + u_{Rj} \\
&\beta_{Mj} = \gamma_{M0} + u_{Mj} \\
&\beta_{Aj} = \gamma_{A0}
\end{aligned}
\qquad (1)
$$

Outcome values were generated based on the following model:

$$
\begin{aligned}
&Y_{ij} = \beta_{0j} + \beta_{Tj} T_{ij} + \beta_{Rj} X_{Rij} + \beta_{Mj} X_{Mij} + \beta_{Aj} X_{Aij} + r_{ij} \\
&\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j} \\
&\beta_{Tj} = \gamma_{T0} + u_{Tj} \\
&\beta_{Rj} = \gamma_{R0} + u_{Rj} \\
&\beta_{Mj} = \gamma_{M0} + u_{Mj} \\
&\beta_{Aj} = \gamma_{A0}
\end{aligned}
\qquad (2)
$$

The simulation manipulated two factors between cells: the ICCs of the student-level covariates and the number of students within each cluster. Within cells, it varied the PS model, matching method, and balance measures. All conditions had a total of 50 clusters. In one condition, the PS model was the correctly specified, data-generating model, and in the other conditions, the PS models were misspecified variations of the model. The matching methods included pooled matching, where a match could be made across clusters; within-cluster matching, where matching could only occur within the same cluster; and two-stage matching, where a match was first attempted within the same cluster but if no close matches were available, it was made outside the cluster (Rickles & Seltzer, 2014). All matching methods used nearest-neighbor 1:1 matching with a caliper of .2 standard deviations of a PS. To assess balance, we varied the type of balance measure (ASB or variance ratio), the summarization of balance across clusters (pooled or within-cluster), and the summarization of balance across covariates (a simple mean or a weighted mean based on the strength of the covariate in predicting the outcome).

The study procedures are summarized in Figure 1.

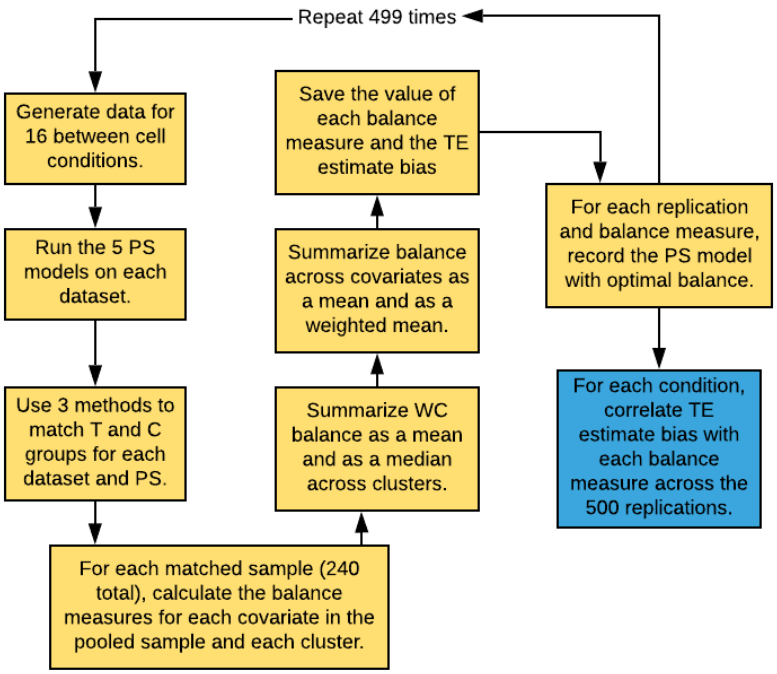

Figure 1. Flow of simulation procedures from data generation to outcome estimation. PS=propensity score; T=treatment; C=control; WC=within-cluster; TE=treatment effect.

Results:

The results indicated that overall across conditions, the ASB pooled across clusters was most highly correlated with bias in the TE estimate (Figure 2). Specifically, the pooled ASB in which the mean ASB across covariates was weighted according to the strength of the covariate's relation to the outcome was most strongly correlated in most conditions, followed by the pooled ASB in which all covariates have equal weight in the mean. Nearly all other balance measures had correlations with TE estimate bias hovering near 0, indicating that they were not as useful for predicting TE estimate bias. In the case of the largest cluster size and within cluster matching, the pooled ASB measures were negatively correlated with TE estimate bias. This is likely because the TE estimate bias was extremely low in this condition, and the pooled ASB is most strongly related to TE estimate bias when there is more variation and greater levels of TE estimate bias (Figure 3).

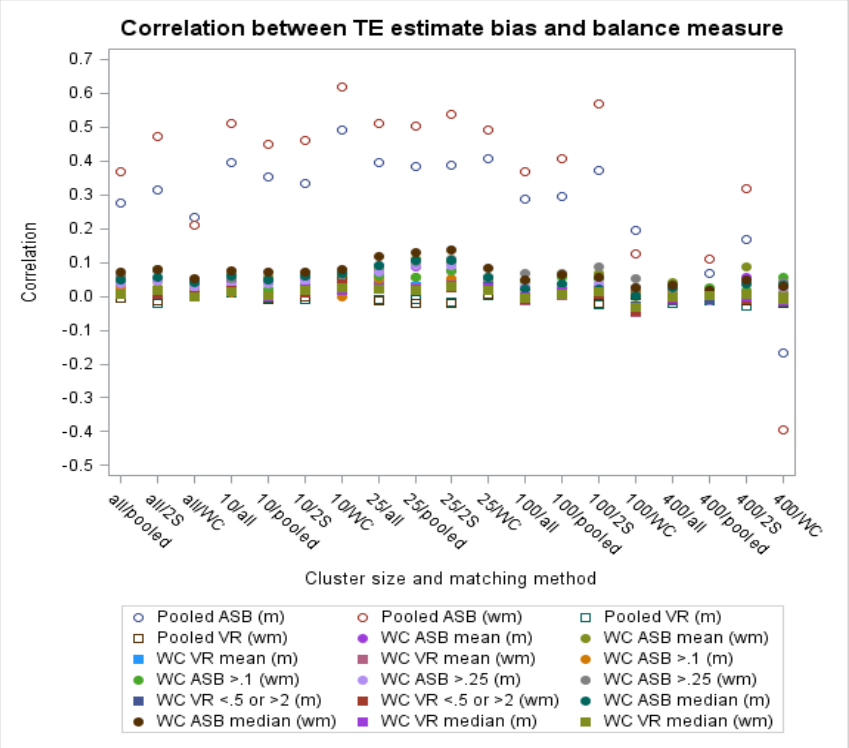

Figure 2. Correlations between treatment effect (TE) estimate bias and balance measures, by cluster size and matching method. ASB=absolute standardized bias; VR=variance ratio; WC=within-cluster balance measure; m=equally weighted mean; wm=weighted mean (according to the covariate's relation to the outcome measure). Values represent the mean correlation across ICCs and PS models.

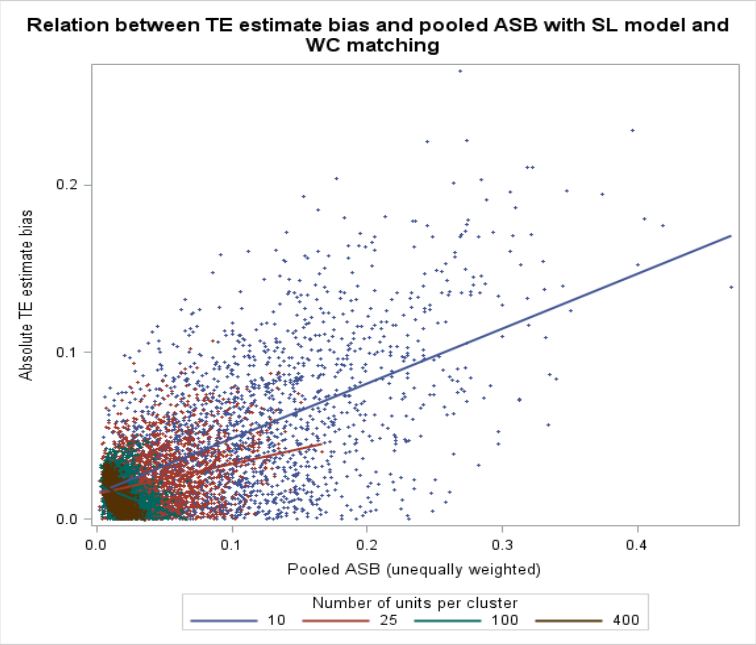

Figure 3. Scatterplot depicting the relation between absolute treatment effect (TE) estimate bias and the pooled, weighted absolute standardized bias balance measure for the conditions with the single-level (SL) propensity score model and within-cluster matching, disaggregated by cluster size.

Conclusions:

In interpreting the results, one must be aware of the conditions that were not tested and the resulting limitations. For example, the study did not consider other types of TE estimation methods that could lead to less biased results, such as separate TEs for each cluster (Kim & Seltzer, 2007) and TEs that include covariate regression adjustment (Robins et al., 1994, Funk et al., 2011). Assessing the correlation between the balance measures and different types of TE estimates will be an important next step.

Nevertheless, the study increased the knowledge of PS methods and balance measures in multilevel settings. The results suggest that in most cases the pooled ASB will have the highest correlation with the TE estimate. The results also suggest that when averaging the balance results across many covariates, researchers should weigh them according to their likely influence on the outcome measure.

References:

Arpino, B., & Cannas, M. (2016). Propensity score matching with clustered data. An application to the estimation of the impact of caesarean section on the Apgar score. *Statistics in Medicine, 35*, 2074-2091.

Arpino, B., & Mealli, F. (2011). The specification of the propensity score in multilevel observational studies. *Computational Statistics & Data Analysis, 55*(4), 1770-1780.

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine, 28,* 3083-3107.

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., & Davidian, M. (2011). Doubly robust estimation of causal effects. *Practice of Epidemiology, 173*(7), 761-767.

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

Rickles, J. H., & Seltzer, M. (2014). A two-stage propensity score matching strategy for treatment effect estimation in a multisite observational study. *Journal of Educational and Behavioral Statistics, 39*(6), 612-636.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of repression coefficients when some regressors are not always observed. *Theory and Methods, 89*(427), 846-866.

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

Thoemmes, F. J., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research, 46*(1), 90-118.

Thoemmes, F. J., & West, S. G. (2011). The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research, 46*(3), 514-543.

Tourangeau, K., Nord, C., Lê, T., Sorongon, A.G., Hagedorn, M.C., Daly, P., & Najarian, M. (2015). *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), User's Manual for the ECLS-K:2011 Kindergarten Data File and Electronic Codebook, Public Version* (NCES 2015-074). U.S. Department of Education. Washington, DC: National Center for Education Statistics.