Propensity Score Analysis with Nested Data: Comparing Single and Multilevel Estimates

Patricia Rodríguez de Gil[1], Rheta E. Lanehart[2], Aarti P. Bellara[1]
Jeffrey D. Kromrey[1], Eun Sook Kim[1], Kathryn M. Borman[2], Reginald Lee[1]
[1] Educational Measurement and Research, University of South Florida, College of Education,4202 E. Fowler Ave. EDU 105, Tampa, FL 33620
[2] Alliance for Applied Research in Education and Anthropology, University of South Florida, 4202 E. Fowler Ave. SOC 107, Tampa, FL 33620

Abstract
Propensity score (PS) methods provide viable strategies for reducing selection bias in nonexperimental (observational) studies. Most research on PS methods model the treatment assignment so that the estimated probability of receiving treatment allows for the identification of comparable individuals based on their individual characteristics. However, in nested data structures selection bias might result not only from differences in the characteristics of the individuals but also from differences in group membership. This study investigated differences in PS results from single-level and multi-level models. Data from an NSF funded project included school transcripts, demographics, enrollment, and achievement data. The impact of special educational programs on advanced mathematics course enrollment was investigated. Data were analyzed by comparing PS distributions, estimating the correlations between the two sets of propensity scores, and comparing the estimates of treatment effects. Results suggest a strong correlation between the PS obtained from single-level and multi-level models and only modest differences in resulting score distributions and estimates of treatment effects.

Key Words: Propensity scores, hierarchical linear modeling, observational studies

1. Background of Propensity Score Analysis

1.1 Overview of Propensity Score Analysis

The propensity score is a relatively new statistic used to reduce bias in observational studies. First introduced by Rosenbaum and Rubin (1983), propensity score analysis attempts to mimic the balance that occurs in randomized experiments. Causal relationships exist between two variables when the following hold true: (a) the cause precedes the effect, (b) the cause is related to the effect, and (c) no plausible alternative explanations for the effect exist other than the cause (Shadish, Cook, & Campbell. 2002). Treatment effects are estimated by a counterfactual model, which is simply the difference between what did happen after an individual received a treatment versus what would have happened if the same individual did not receive the treatment (Campbell & Stanley, 1963; Holland, 1986; Rubin, 2010; Shadish et al., 2002). Theoretically, an exact effect would be measured by taking the difference an individual received on both treatments; however, it is not plausible to assign the same individual to both treatment and control groups. This impossibility is often referred to as the Fundamental Problem of Causal Inference (Holland, 1986, p. 947). Therefore, treatment effects should be estimated by a counterfactual model using propensity score analysis. The process of randomization guarantees the two groups, on average, will be balanced on all characteristics at the beginning of the experiment, and thus able to yield estimates of the average treatment effect (ATE). In contrast, experiments which do not employ random assignment techniques, yet aim to explore causation, provide “less compelling support for counterfactual inferences” (Shadish, et al., 2002, p. 14) because groups are not probabilistically similar. In addition, causal relationships from non-manipulable variables may also be identified.

Propensity scores predict an individual's probability for being assigned to the treatment group, thus ranges from 0 to 1. The closer the individual’s PS is to 1, the stronger the prediction for being in the treatment group; conversely, the closer the score is to 0, the stronger the prediction for being in the comparison group. When units from the treatment and control group have the same propensity score, it is assumed that the probability of being assigned to the treatment group is the same for each of these individual units, conditional upon the observed covariates. When there is no overlap in PSs between the groups, it is believed that unobserved covariate(s) are accounting for the difference in groups (Stuart, 2010).

1.2 Single Level vs. Multilevel Analysis

The majority of the research on propensity score analysis has focused on research designs where the strongly ignorable treatment assignment assumption has been violated but the stable unit treatment value assumption (SUTVA) is assumed to be satisfied. The strongly ignorable treatment assignment assumption requires the assignment to condition be independent and not associated with the outcome or other factors; hence studies that are not using a random assignment process are the focus. SUTVA is defined as an "a priori assumption that the value of Y for unit *u* when exposed to treatment *t* will be the same no matter what mechanism is used to assign treatment *t* to unit *u* and no matter what treatments the other units receive" (Rubin, 1986, p. 961). Simply, SUTVA assumes the outcomes from two individuals, irrespective of treatment assignment, are independent from one another. Limited research has focused on propensity score analysis when both of the aforementioned assumptions have been violated (e.g., observational studies in which outcomes from individuals are not independent of each other). Often in educational settings, certain schools, demographic areas, or neighborhoods have differing student achievement levels, and arguably certain settings are predisposed to offer advantageous learning environments over others (Oakes, 2004). Additionally, assignment to condition may be a result of the contextual factors of the cluster. This is where causal inference is complicated as each single unit’s potential outcome is not only dependent on treatment assignment but also on cluster membership and any cluster level contextual factors —a violation of SUTVA (Thoemmes, 2009).

Multi-level modeling (MLM) is a family of statistical analyses used to evaluate nested data (Raudenbush & Bryk, 2002). Multi-level models improve the estimation of individual effects in nested data by accounting for the dependencies among the units, adjusting the standard error properly, and partitioning the variance at all levels (Raudenbush & Bryk, 2002). Additionally, MLM allows for cross-level interactions, which explain how variables measured at one level affect the associations occurring at another level (Raudenbush & Bryk, 2002). One way to resolve the challenges associated with multilevel data would be to consider the causal inference at the cluster level, where the clusters are treated as individual units. However, in order to draw inferences on individuals, statistical adjustments using MLM need to be applied in order to account for the complexity of both the research design and structure of the data.

2. Purpose

The purpose of this study was to compare the sample propensity scores obtained from single-level and multi-level models applied to educational data characterized by a natural hierarchical structure (i.e., students nested within schools). Direct comparisons were made of the distributions of propensity scores in the samples, the correlations between the single-level and multi-level propensity scores were computed, and the estimated treatment effects obtained from the outcome models using the two sets of propensity scores were compared.

3. Method

3.1 Data Sources

Longitudinal data (2002-2007) containing student high school transcript records, student demographics, achievement scores from standardized tests, and student school enrollment were obtained from the Department of Education in a southeastern state through a National Science Foundation grant designed to study STEM experiences in high school career academies. Career academies (CA) are small learning communities within a larger high school whose participants are grouped as grade-level cohorts. The CA cohorts move through a sequence of courses taught by the same interdisciplinary team of teachers, each teaching either a career and technical education (CTE) course centered on the career academy's theme or a core academic course. The ultimate goal of the career academy model is to provide a rigorous and relevant curriculum, enabling students to enter the workforce and/or any level of post-secondary education directly following high school.

3.1.1 Sample

The sample consisted of students (*n* = 134,597) who were enrolled in coursework in regular high schools (*n* = 343) during the 2006-07 school year. Certain exceptional students and students not enrolled in regular high schools were deleted from the sample.

3.2 Propensity Score Method

3.2.1 Multiple Imputation of Missing Data

Variables in the dataset were evaluated for missing values. To address the problem of missingness, multiple imputations were performed using PROC MI in SAS® 9.2 (SAS Institute, 2010) using the default value, which created five datasets with imputed values for variables with incomplete data. The distribution equivalence of the imputed datasets were compared using the Kolmogorov-Smirnoff test

3.2.2 Selection of Covariates

We reviewed the extant literature to determine the relevant variables to include as covariates in generating propensity scores. We chose variables related to career academy participation and rigorous high school courses.

We included covariates related to students' demographic and home background such as gender, race, home language, free or reduced price lunch status, and migrant status. Students' school related covariates were also included. All student-level covariates are binary except two ordinal measures of math and reading achievement levels. At the school level, eight covariates were considered. The selected school-level covariates are all continuous and include school average math and reading scaled scores, the number of full time equivalent teachers, student/teacher ratio, and total school-level demographic covariates. In total, 23 covariates were selected for the propensity score estimation (Table 1).

Table 1: Descriptive statistics of covariates

Variable	Type	Mean/Proportion (SD)	Minimum	Maximum
Student level				
Female	Binary	.519 (.500)	0	1
Race				
Asian	Binary	.030 (.170)	0	1
Black	Binary	.225 (.418)	0	1
Hispanic	Binary	.250 (.433)	0	1
Home language				
English	Binary	.720 (.449)	0	1
Spanish	Binary	.182 (.385)	0	1
Other Language	Binary	.099 (.298)	0	1
Free/ Reduced Lunch	Binary	.288 (.453)	0	1
Migrant	Binary	.004 (.062)	0	1
Exceptional Student Education	Binary	.081 (.273)	0	1
Gifted	Binary	.056 (.230)	0	1
Specific Learning Disability	Binary	.069 (.253)	0	1
Limited English Proficient	Binary	.197 (.398)	0	1
Math Achievement Level	Ordinal	2.925 (1.209)	1	5
Reading Achievement Level	Ordinal	2.631 (1.099)	1	5
School level				
Total 8th grade N	Continuous	441.032 (145.116)	2	810
Total school- level ethnicity	Continuous	1314.420 (406.468)	21	2490
Total school- level free lunch	Continuous	479.981 (324.054)	0	1522
Total school- level reduced price lunch	Continuous	123.708 (61.860)	0	289
Full time equivalent teachers	Continuous	63.472 (17.582)	0	109
Student/teacher ratio	Continuous	20.503 (4.293)	1	90
Math Scaled Score	Continuous	1902.290 (190.568)	1025	2605
Reading Scaled Score	Continuous	1893.790 (252.407)	886	2790

3.2.3 Balance Diagnostics

To assess balance between the groups on the selected covariates, we examined box plots and computed Cohen's *d* (standardized mean difference) for each continuous variable. A standardized mean difference smaller in magnitude than 0.25 (Stuart, 2010) was used as our criterion for acceptable balance. Balance for dichotomous and discrete ordinal covariates was evaluated using odds ratios.

3.3 Propensity Score Model

Propensity scores were used to reduce selection bias in our observable data and provide a quasi-experimental framework to compare average treatment effect between groups (Rubin, 1973; Rosenbaum & Rubin, 1983; Stuart, 2010). Incoming 9th grade students were matched on 8th grade demographics, standardized test scores, and school covariates using PROC LOGISTIC (single level analysis) and PROC GLIMMIX (multi-level analysis) in SAS® 9.2 (SAS Institute, 2010). Posterior probabilities (i.e., propensity scores) were obtained for enrolling in a STEM career academy. The propensity scores represent the estimated probability that a student would be participating in a STEM career academy based on that student's covariate values.

The distributions of the estimated propensity scores were evaluated by common support. Areas of non-overlap in the propensity score distributions were trimmed by discarding cases in the region of non-overlap. Assessment of the equivalence of the propensity score distributions included a Kolmogorov-Smirnoff test, a comparison of summary statistics from each distribution, and graphical displays.

3.4 Stratification on the Propensity Score

The propensity scores from each model were stratified into five subclasses using the quintiles and assigned ranks from 1 to 5 (Rosenbaum & Rubin, 1984, 1985; Austin, 2007; Stuart, 2010). Propensity score stratification was completed for each imputation resulting in five samples of cases stratified on propensity scores. The number of observations in each stratum for the single level and multilevel PS analyses for the first imputation are presented in Table 2.

3.4.1 Assessment of Balance across Strata

Balance was assessed within each stratum, then pooled across strata to obtain an overall balance statistic for each covariate (Harder, Stuart, & Anthony, 2006). Multilevel linear models were used for assessing balance on continuous covariates and multilevel generalized linear models were used for dichotomous and discrete ordinal covariates.

Table 2: Total Number of Individuals in Treatment and Control group by Stratum for Single Level and Multilevel PS for Imputation 1

	Single Level PS			Multilevel PS		
Stratum	Control	Treatment	Total	Control	Treatment	Total
1	26725	124	26849	26474	111	26585
2	26660	189	26849	26404	182	26586
3	26563	286	26849	26339	246	26585
4	26410	439	26849	26134	452	26586
5	25986	863	26849	25679	906	26585
Overall	132344	1901	134245	13103	1897	132927

3.5 Estimation of the Effect of STEM Career Academy on Calculus Course Enrollment

To estimate the effect of STEM Career Academy on enrollment in calculus during high school, multilevel models with students nested within schools were run for the five imputations using Proc GLIMMIX in SAS 9.2 (SAS Institute, 2010). Odds ratios and confidence intervals were obtained for each stratum, as well as an overall effect estimate pooled across strata.

4. Results

4.1 Propensity Score Comparisons

A strong, positive correlation (*r* = 0.67) was demonstrated between the single level and multi-level propensity scores (Figure 1).

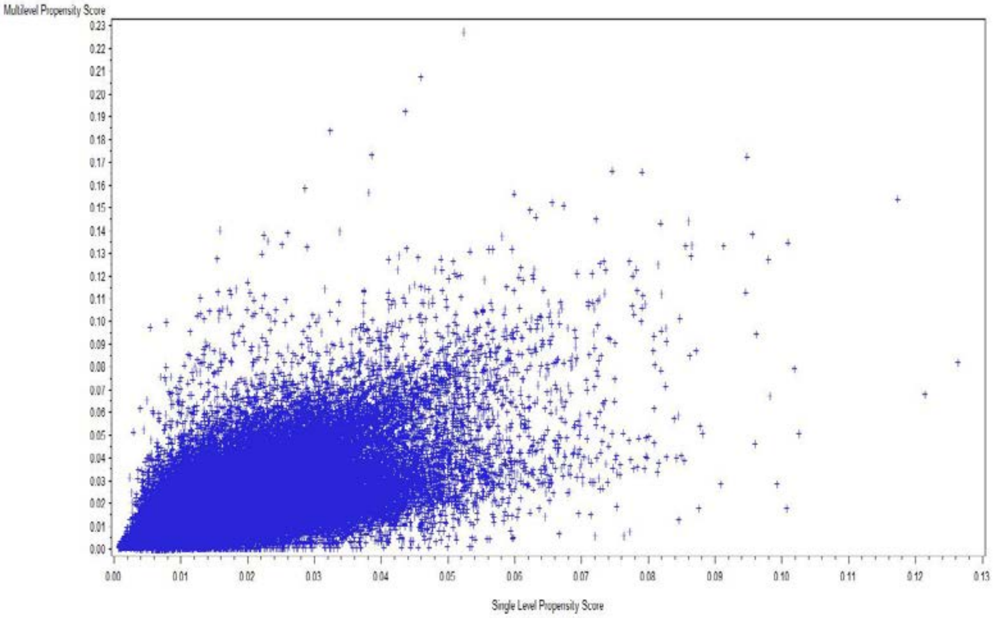
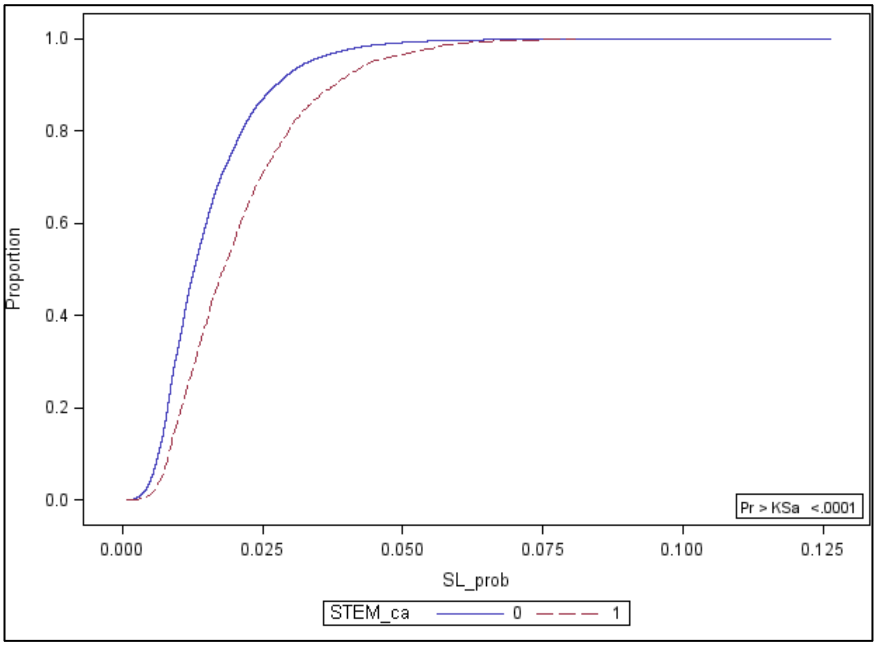
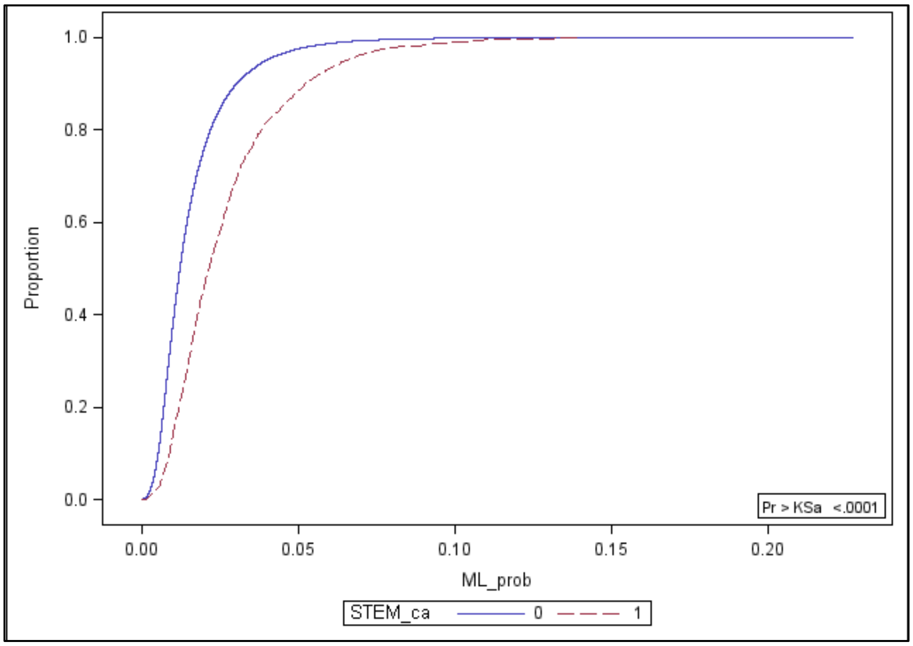


Figure 1: Bivariate plot of single level and multilevel propensity scores

Comparison of the Kolmogorov-Smirnoff distributions of propensity scores for STEM career academy students and non-STEM career academy students (Figures 2 and 3) indicated very little difference between the two groups for propensity score distributions estimated from single and multi-level models both before trimming (single level: *D*= 0.223, p< 0.001; multi-level: *D*= 0.326, p< 0.001) and after trimming (single level *D*= 0.221, p< 0.001; multi-level: *D*= 0.324, p< 0.001).

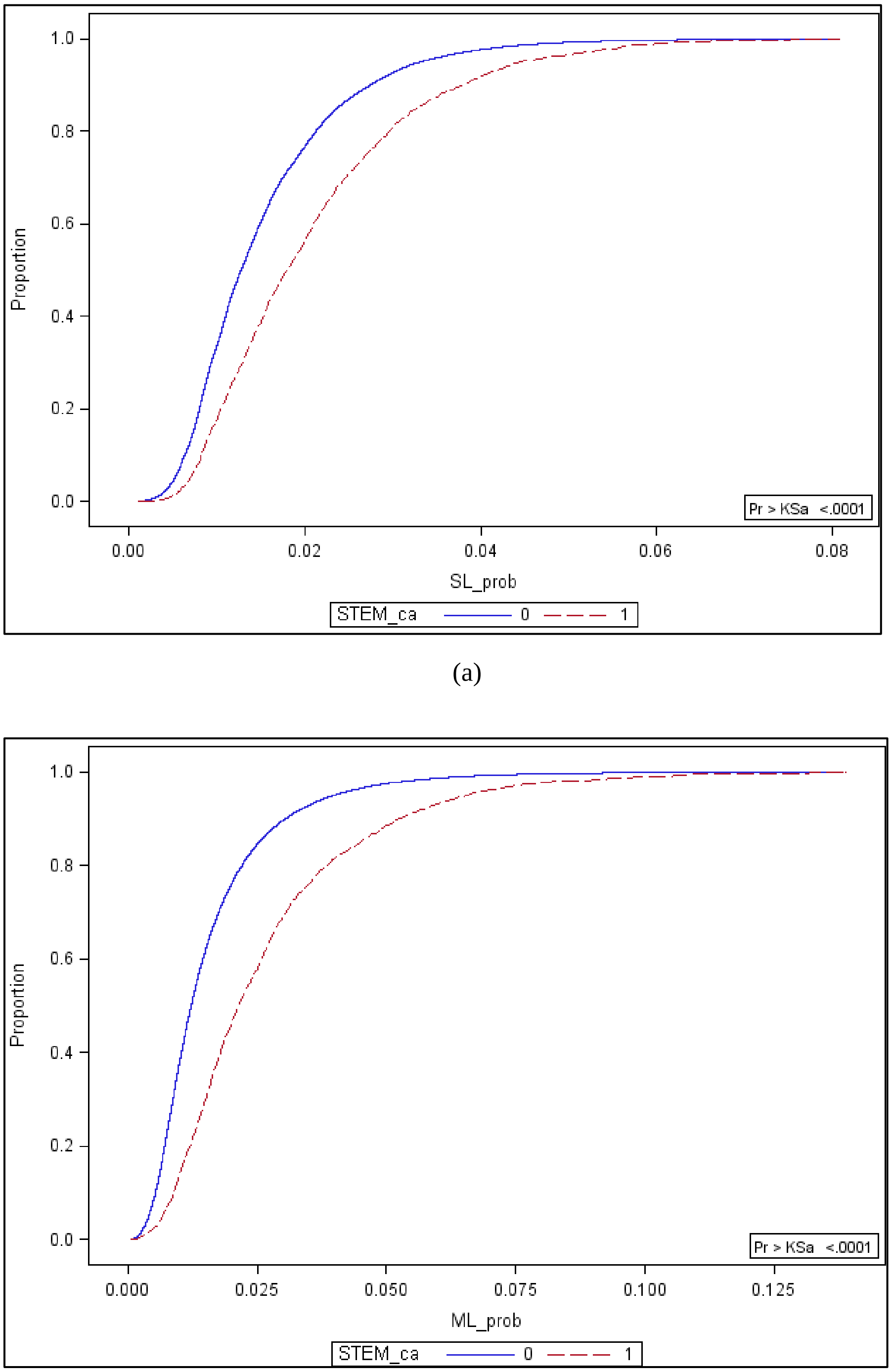


(a)



(b)

Figure 2: Distribution of untrimmed propensity scores for STEM Career Academy groups estimated with single level (a) and multi-level (b) models.

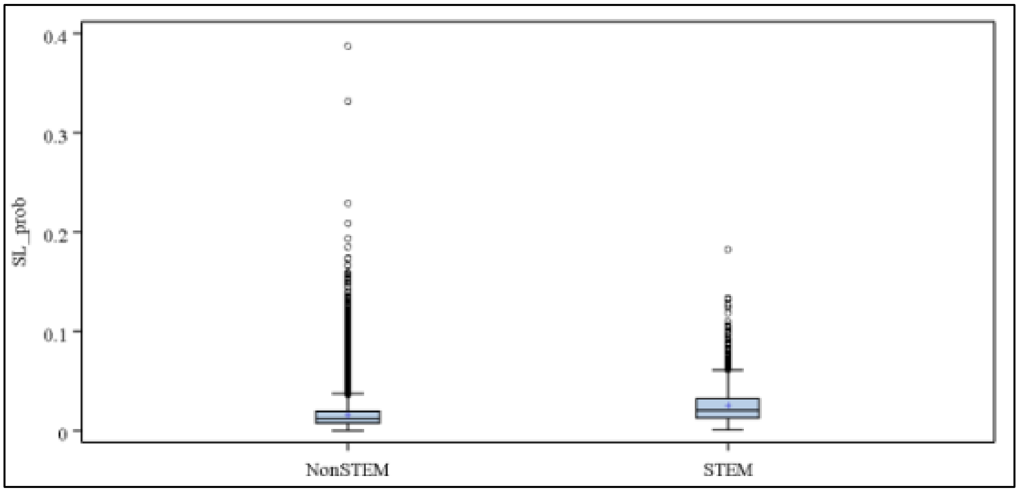


(a)

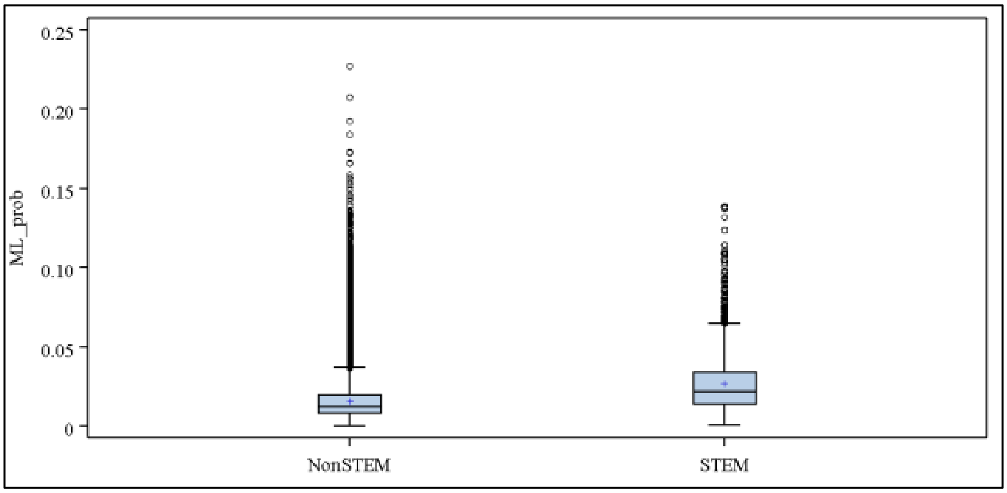
(b)

Figure 3: Distribution of trimmed propensity scores for STEM Career Academy groups estimated with single (a) and multi-level models (b).

The boxplots of single and multi-level propensity scores before and after trimming indicated a wider range of scores estimated in the single level models (Figures 4 and 5). The mean propensity score for the STEM group was slightly higher than the non-STEM group for all models.

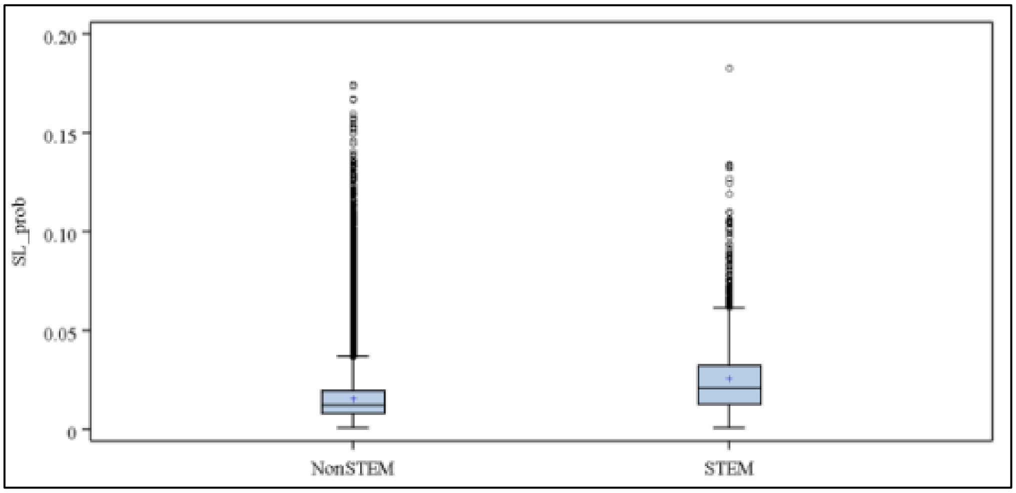


(a)

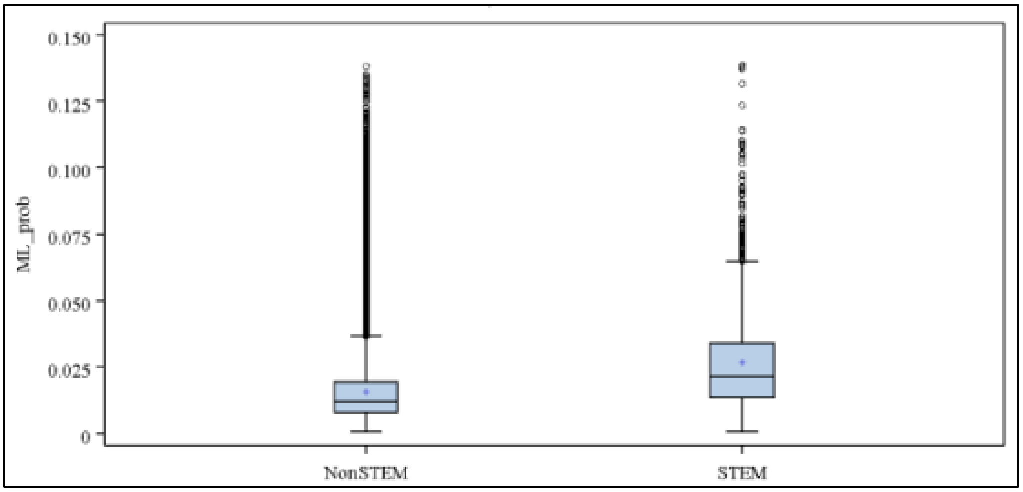


(b)

Figure 4: Boxplots of single level (a) and multi-level (b) propensity scores before trimming



(a)



(b)

Figure 5: Boxplots of single level (a) and multi-level (b) propensity scores after trimming

4.2 Balance of Covariates

An analysis of the covariate balance is presented in Table 3. The balance statistics were obtained by fitting multilevel linear models (for continuous covariates) and multilevel generalized linear models (for binary and discrete ordinal covariates) to assess treatment group differences in the covariates after stratification on the PS. Analogous models without stratification provided balance statistics before conditioning on the PS.

The balance estimates from the single level and multilevel PS models were similar across most covariates. For the binary and discrete ordinal covariates, only “Other Language” evidenced an odds ratio greater than 2.0 after conditioning (a result that was consistent across the two models). For the continuous covariates, all standardized mean differences were less the 0.25 in absolute value except for the students’ 8th grade Math Scaled Score after conditioning on the PS from the single level model (*d* = -0.258).

Table 3: Covariate Balance After Conditioning on Single Level and Multilevel Propensity Scores

	Before	After Conditioning	
Covariate	Conditioning	Single Level	Multilevel
Binary			
Female	1.014	1.145	1.204
Asian	0.601	0.952	0.919
Black	0.968	1.265	1.311
Hispanic	1.333	1.499	1.508
English	0.728	0.617	0.620
Spanish	1.017	1.133	1.160
Other Language	1.602	2.327	2.125
Free/ Reduced Lunch	1.051	1.270	1.287
Migrant	2.961	0.978*	0.786*
Exceptional Student Education	2.608	1.170	1.502
Gifted	0.657	0.732	0.770
Specific Learning Disability	2.692	1.912	1.507
Limited English Proficient	1.098	1.413	1.385
Discrete Ordinal			
Math Achievement Level	0.440	0.569	0.666
Reading Achievement Level	0.506	0.694	0.749
Continuous			
Total 8th grade N	-0.007	0.047	0.050
Total school-level ethnicity	-0.010	0.042	0.046
Total school-level free lunch	-0.004	0.003	0.202
Total school-level reduced price lunch	0.180	0.063	0.064
Full time equivalent teachers	-0.042	0.044	0.098
Student/teacher ratio	-0.005	0.010	0.016
Math Scaled Score	-0.425	-0.258	-0.212
Reading Scaled Score	-0.362	-0.211	-0.170

Note. Odds Ratios are reported for binary and ordinal variables while standardized mean differences are reported for continuous variables.
* = Model nonconvergence

For several covariates, the balance between the STEM Career Academy group and the control group was superior prior to conditioning on the PS. Such a result suggests that the functional form of the PS models (linear, additive models) may have been incorrect. Investigation of more complex propensity score models (i.e., including quadratic terms and interactions among the covariates) would be recommended, but we needed to maintain identical models for both the single level and multilevel PS so that such model differences would not confound our comparisons of the two sets of PS.

4.3 Estimates of Treatment Effect

The dependent variable of interest in this study was whether or not students enrolled in a calculus course; therefore, estimates of treatment effects were analyzed using hierarchical logistic regression models (Table 4). Regression coefficients were estimated within each stratum and pooled across strata to obtain an overall estimate for the entire sample. Confidence intervals were constructed for each stratum and pooled across strata to determine whether there were underlying differences in outcome for a subsample of individuals. The patterns within each PS method (single level and multilevel) as well as across strata were all similar, suggesting that those who enrolled in a STEM career academy had greater odds of enrolling in an advanced calculus course.

Table 4: Treatment Effect Estimates by Stratum for Single Level and Multilevel Propensity Score Analyses

	Single Level PS				Multilevel PS			
			95% CI				95% CI	
Stratum	β	OR	LL	UL	β	OR	LL	UL
1	0.65	1.91	0.26	13.84	0.71	2.04	0.47	8.84
2	0.63	1.87	0.83	4.23	0.13	1.14	0.50	2.62
3	0.67	1.95	1.19	3.19	0.52	1.68	0.99	2.88
4	0.65	1.92	1.28	2.87	0.71	2.03	1.40	2.94
5	0.68	1.97	1.53	2.53	0.64	1.90	1.46	2.46
Overall	0.68	1.97	1.66	2.33	0.63	1.87	1.58	2.22

5. Conclusions

The results of this comparison between propensity scores estimated from a single level logistic regression model and those estimated from a multilevel generalized linear model suggest few differences between the two sets of propensity scores. The propensity scores themselves evidenced a substantial degree of correlation and the resulting score distributions were similar. Further, few differences were evident in the extent to which stratification on the resulting scores produced balance between the two groups on the set of covariates. Finally, the estimates of treatment effects obtained after stratifying on the single level and multilevel PS were similar, both in the magnitudes of the point estimates of odds ratio and in the widths of the confidence intervals.

These results, of course, are based on a single large sample of actual field data and the similarities between the two sets of propensity scores could result from idiosyncrasies in these data. More thorough investigations of similarities and differences between propensity scores obtained from single level and multilevel models should be conducted under the controlled conditions of simulation research.

Acknowledgements

Work on this paper was supported for the first two and last authors by a grant provided by the National Science Foundation (NSF ITEST #0833503 awarded to Kathryn M. Borman, P.I.).

References

Austin, P. C. (2007). The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in Medicine, 26,* 3078-3094.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and Quasi-experimental Designs for Research.* Chicago: Rand McNally.

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.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81,* 945-960.

Oakes, J. M. (2004). The (mis)estimation of neighborhood effects: Causal inference for a practicable social epidemiology. *Social Science and Medicine, 58,* 1929-1952.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods.* Thousand Oaks: Sage Publications.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika,* 70, 41-45.

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.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician,* 39, 33-38.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics,* 29, 159-184.

Rubin, D. B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association, 81*(396), 961-962.

Rubin, D. B. (2010). On the limitations of comparative effectiveness. *Statistics in Medicine, 29*(19), 1991-1995.

SAS Institute (2010). *SAS/STAT® Users Guide 9.22,* (Cary, N. C., SAS Institute, Inc.).

Shadish, W. R. (2002). Revisiting field experimentation: Field notes for the future. *Psychological Methods, 7*(1), 3-18.

Shadish, W. R., Cook T. D., & Campbell, D. T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference.* Belmont, CA: Wadsworth, Cengage Learning.

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

Thoemmes, F. J. (2009). The use of propensity score with clustered data: A simulation study. Unpublished Dissertation. AAT: 3380671.