A Propensity Score Matching with Multiple Imputation Approach: Studying the Effects of Remedial Math-Taking in College on Degree Attainment

Meghan Ann Clovis
Lecturer
Department of Mathematics
College of Arts & Sciences
University of Miami
1365 Memorial Drive, Ungar 515, Coral Gables, FL 33146

Mido Chang* Corresponding author
Professor of Research Methodology
Department of Counseling, Recreation, & School Psychology
Florida International University
11200 SW 8th St. ZEB 245A, Miami, FL 33199

Abstract

The purpose of this study was to investigate whether multiple imputation procedures (MI) and propensity score matching (PSM) would improve the estimation of remedial math-taking on college-degree attainment. The primary motivation was to explore a robust research method in an area of educational research that suffers from the inability to conduct a randomized controlled design. Analyses were conducted on both matched and unmatched groups, as well as on 12 multiply imputed complete-case data sets. Logistic regression analyses, both with and without the inclusion of covariates, indicated that remedial math-taking in college was a significant negative predictor of degree attainment. These results were consistent across unmatched groups, matched groups, and all 12 multiply imputed data sets. Neither multiple imputation procedures nor propensity score matching resulted in a significant improvement in the ability to control for preexisting group differences.

Keywords: degree attainment, remedial math, propensity score matching, multiple imputation

Introduction

Standard statistical analysis methods are not designed for comparing nonequivalent groups, such as those that often exist in educational research. Though a randomized controlled trial is often considered the gold standard in research, non-random assignment of participants to treatment and control groups, as well as selection bias, are common problems in educational research (Padgett et al., 2010; Shadish & Steiner, 2010; Titus, 2007). When students are not randomly assigned to treatment and control groups, comparison groups cannot be assumed to be equivalent on any covariate measures that may affect the outcome under investigation because students in each group may be systematically different.

One method that has been suggested to address the problem of comparing nonequivalent groups is propensity score matching (PSM). Supporters of PSM contend that it can mimic a randomized controlled trial by removing imbalance in covariate measures between comparison groups (Austin, 2011; Hill, 2004; Mitra & Reiter, 2016; Rosenbaum & Rubin, 1983). The use of PSM in educational research is growing (Byun et al., 2015; Clark & Cundiff, 2011; Fan & Nowell, 2011; Giani et al., 2014; Henderson & Chatfield, 2011; Luellen et al., 2005; Melguizo et al., 2011; Padgett et al., 2010; Titus, 2007; Vaughan et al., 2014). PSM involves forming subgroups of matched sets of treated and untreated participants from the larger sample using a propensity score. In a non-randomized study, the propensity scores are estimated using the covariate measures in the sample and are considered similar if the difference is within a predefined range (Shadish & Steiner, 2010). Logistic regression is a common method for estimating the propensity scores when the dependent variable is dichotomous and the method of analysis of treatment effects is logistic regression (Holmes, 2014).

Because the estimation of propensity scores relies on complete case analysis for all covariates used in the estimation process, PSM results in listwise deletion of cases with missing data. As such, PSM is often combined with missing data analysis procedures (Hill, 2004; Mitra & Reiter, 2016). When data are missing, it is important to consider the cause of the missing data prior to analysis. The researcher must investigate the patterns in, and impact of, the missing data in the analysis.

One recommended method to handle missing data is multiple imputation (Carlin, Greenwood, & Coffey, 2003; Harel & Zhou, 2007; Miles, 2016; Reist & Larsen, 2012). Multiple imputation (MI) uses multiple variables to estimate missing data, creates multiple models for each estimate, and then combines the results of these models (Holmes, 2014). After MI, analyses are conducted on the pooled imputed data sets. Theoretically, the more models used to create the pooled estimates, the more valid the statistical analysis will be (Harel, 2007; Yuan, 2010). There is no consensus on the number of imputations one should use, but more recent studies suggest that the number of imputations should be greater than or equal to the largest percentage of missing values (White et al., 2011).

MI is appealing because it results in estimates of missing values and maintains the original sample size, but it is imperative that researchers investigate differences in the outcomes of statistical analyses both with and without the imputed values given the untestable assumption about the missing data mechanism (Meyers et al., 2013). Statistical packages that function to create MI data sets do not provide much flexibility in working with the combined results from MI. In many cases, this is because there are not statistically valid methods for combining common statistics, such as those obtained in regression analyses (Miles, 2016; Mood, 2010). Although many studies cite "Rubin's rule" for combining estimates, statisticians disagree on whether these estimates can be combined or, more importantly, interpreted so easily in the context of logistic regression (Mood, 2010).

PSM has the potential to make comparison groups more equivalent, thereby improving the foundation for making causal inferences. Several researchers have conducted extensive investigations into the claims that PSM can mimic a randomized experiment. Peikes et al., (2008) cited major limitations in PSM based on their investigation, emphasizing that it requires correct covariate selection and a very large sample size. PSM is labor-intensive and time-consuming. Furthermore, there is no way to determine in advance if PSM will work. Hill et al., (2011) conducted a similar investigation using longitudinal data and also cited a number of limitations to PSM, cautioning that effective analysis requires a lot of choices be made by the researcher including model fitting, matching method, deciding if/when groups are balanced enough, and how to analyze the results of PSM. Rosenbaum and Rubin (1985), the major contributors to the theory of PSM stated, "matching, when successful, is a persuasive method of adjusting for imbalances in observed covariates" (p. 33). In other words, if PSM works, it works well, but there is no way of knowing whether it will work, whether it accomplished the goal of creating statistically equivalent comparison groups, or if it produced unbiased estimates of treatment effects.

We chose to focus our investigation of PSM and MI on an area of educational research that necessarily suffers from the inability to conduct a randomized controlled trial–the effects of remedial-math taking on degree attainment for college students. The success (or lack thereof) of students who enter colleges and universities academically underprepared is a topic of interest and great debate among policymakers, colleges/universities, educators and researchers (Bahr, 2010; Bettinger & Long, 2005; Illich et al., 2004; Pretlow & Washington, 2011). Despite the prevalence of math remediation, many students do not successfully complete their coursework (Attewell et al., 2006; Bahr, 2010). The lack of success of remedial math students has prompted numerous revisions to remedial courses and these revisions and have had inconsistent results (Bahr, 2007; Bettinger & Long, 2005; Bonham & Boylan, 2012; Illich et al., 2004).

As Bettinger and Long (2005) pointed out, "Despite the growing numbers of underprepared students who enroll in remedial courses at community colleges each year, little is known about the causal effects of remediation on student outcomes" (p.17). Students taking remedial courses may be systematically different from non-remedial students and preexisting differences may impact their success. This study was undertaken to investigate the effect of remedial math-taking on degree attainment using PSM and MI.

Methods

The data used in this study were obtained from the public use data file of the Educational Longitudinal Study of 2002 (ELS). ELS was a nationally representative, longitudinal study of high school students. There were six major data collection waves beginning with high school data on sophomores and concluding with postsecondary transcript data in the final wave (Bozick et al., 2007).

Sample and Variables

This study focused on the effect of remedial math-taking in college on degree attainment (the dependent variable) using binary logistic regression, PSM, and MI. The variables used to select a subsample of participants from the ELS data were the known postsecondary institution (PSI) attendance and known remedial math-taking in college. After filtering and removing missing cases on the dependent variable, the working sample size was 10,736.

Nineteen covariates were included in the current study to examine (1) whether preexisting differences predicted participation in remedial math and (2) the effects of participation on degree attainment. Common student demographic variables included were sex, race, socioeconomic status (SES), family composition, working status during senior year in high school, and native language. High school-level variables included school control, school urbanicity, highest math course taken, program concentration, college planning program participation, base-year math proficiency, and base-year reading proficiency.

Postsecondary-level variables included school sector, PSI level combination attended, timing of PSI enrollment, and student loan indicators. Two additional composite variables, which used combinations of high school and postsecondary-level data, were included: postsecondary education pipeline completion and high school attainment indicator. Nominal variables with more than two levels were recoded using dummy variables to create dichotomous variables, resulting in a total of 33 covariate measures.

Our analysis was a multi-stage process. Stage one included preliminary analyses on the sample to investigate preexisting differences on covariate measures between students who did or did not take remedial math in college. We then conducted binary logistic regression with all covariates predicting remedial math-taking to determine if the covariates were significant predictors of our main independent variable. Logistic regression analyses predicting degree attainment was run using two models: (1) remedial math-taking as the sole predictor and (2) remedial math plus all covariates as predictors.

In stage two, we conducted PSM to create a subsample of matched remedial math groups, including all covariates during the matching procedure. We investigated the balance of the matched groups on the covariates and then repeated the analyses from stage one using the matched sample to investigate how PSM performed. In stage three, we analyzed missing data for all covariates and then conducted MI procedures to replace missing values on covariates, obtaining 12 multiply imputed complete-case data sets. After conducting MI, we repeated the analyses conducted in stage one. Finally, in stage four, we conducted PSM on each of the 12 multiply imputed data sets obtained in stage three, forming 12 sets of matched groups. We then repeated the analyses conducted in stage one. The outcomes from all four stages were compared to investigate any differences in the resulting estimates of treatment effects.

Results

Our primary goal in this study was to examine the effect of remedial math-taking on degree attainment. Results are presented in two main sections: analyses of remedial group differences and analyses of the effect of remedial math-taking on degree attainment. In the first section, we present a comparison of group differences using four samples: original data, original matched groups, 12 multiply imputed data sets, and 12 matched MI groups. Also presented are the results of a missing value analysis conducted prior to imputation. The second section presents a comparison of the effects of remedial math-taking on degree attainment, with and without inclusion of the covariates, on the same four samples.

Analyses of Group Differences

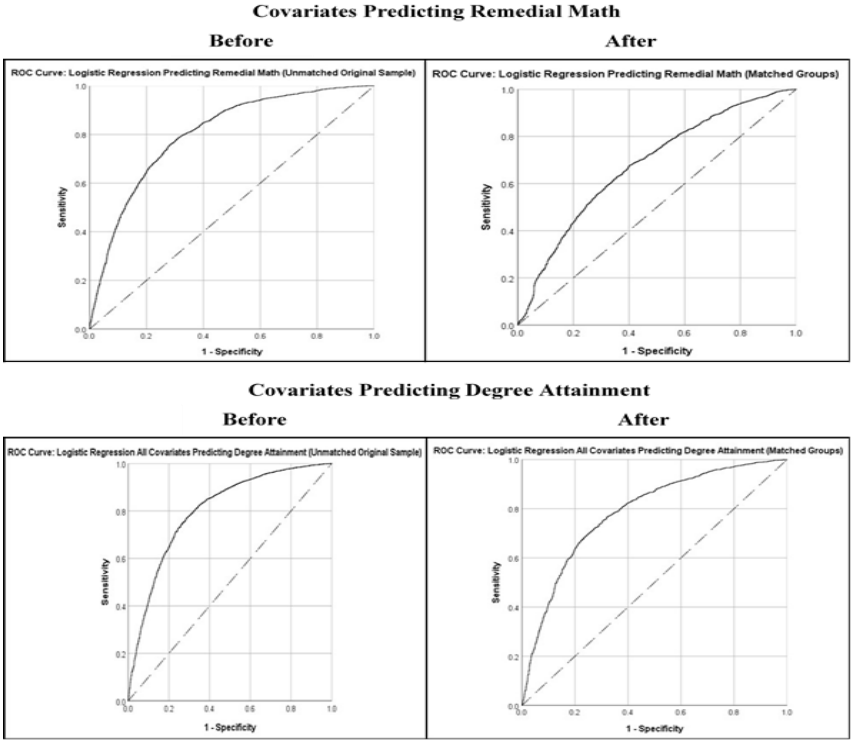
Analyses on differences in remedial math groups included descriptive statistics, *t*-tests, contingency table Chi-square tests of association, standardized differences (effect size) and binary logistic regression predicting remedial group membership. Public school control, general high school program, students of White race, vocational high school program, and 4-year transfer PSI level were used as reference categories in all regression analyses in this study. Standardized residuals were examined for Chi-square tests of association but are not reported in detail to conserve space. In evaluating standardized differences, there is no consensus as to what effect size represents nonnegligible imbalance between groups (Austin, 2009). Holmes (2014) suggests using .20, while Normand et al. (2001) suggest .10. As such, we refrained from making definitive assertions as to whether any imbalance found was negligible, and instead focused on if and how imbalance was affected by PSM and MI. Table 1 summarizes the logistic regression coefficients, standard errors, and odds ratios for the covariates predicting remedial math group membership.

Original Sample

In the initial sample, we found significant differences between remedial groups on all three continuous measures as well as on 23 out of 30 nominal measures. Although statistically significant differences in means were found, this was not entirely unexpected due to the large sample size. Effect sizes were relatively large for continuous, and small for nominal, measures.

A binary logistic regression analysis predicting remedial math-taking (using no remedial math as the reference category) indicated that the model including all covariates provided a statistically significant prediction of remedial math-taking, $\chi^2\left(28, N = 7109\right) = 1853.55$, $p < .001$. The predictive accuracy of the model was 76%. Figure 1 presents a graph of the receiver operating characteristic (ROC) curve, which plots the true-positive (sensitivity) against the false-positive (1 - specificity) rates obtained from the regression analysis. The area under the ROC curve (AUC) is a measure of the discriminating ability of the model–the ability to correctly classify students who did and did not take remedial math. An AUC of .50 is equivalent to tossing a coin (a 50-50 chance of correct classification). The AUC was .80, which is considered a fair discriminating ability. The results of the preliminary analyses indicated there were significant preexisting differences in remedial math groups in the initial sample, which significantly predicted remedial group membership. The group differences had the potential to bias regression estimates of the effect of remedial math-taking on degree attainment.

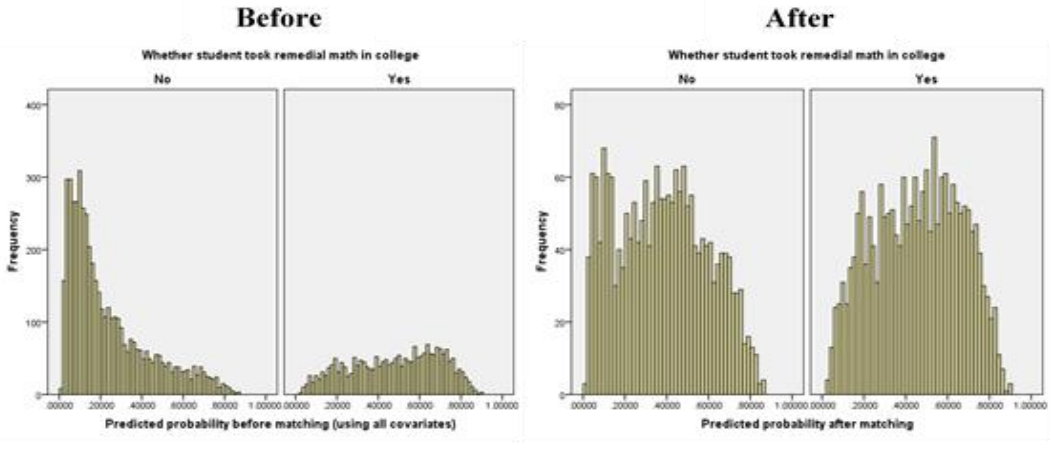
Figure 1. ROC curves: Covariates predicting remedial math-taking and degree attainment before and after matching.



Propensity Score Matching on the Original Sample

Given the potential for biased coefficient estimates due to nonrandom group assignment and preexisting group differences, PSM was conducted on the original sample including all covariates in the model. Due to missing values on the covariates, the initial samples size available for matching was reduced by 34% because of listwise deletion. One-to-one matching with a match tolerance of .1 resulted in a sample size of 3,978 divided equally between remedial groups (representing 44% of the sample available for matching). The predicted probability of group membership after matching is equivalent to the propensity scores used for the matching process. Figure 2 shows side-by-side histograms of the predicted probability of group membership before and after matching. As can be seen from the histograms, the remedial math groups were more similar after matching.

Figure 2. Predicted probability of remedial math group membership before and after matching.



Post-Matching Analyses of Remedial Group Differences

After matching, we found significant mean differences between remedial groups on all continuous variables as well as on 19 out of 30 nominal measures. A comparison of the mean difference and effect size for each continuous measure before and after matching reveals that the mean differences and effect sizes were reduced an average of 75% after matching. However, the standard errors increased an average of 54%. The increase in standard errors may have been the result in the substantial decrease in sample size that resulted from matching.

A comparison of the standardized residuals showed that although the groups differed on the levels of the nominal measures, there were equal residuals (for each level of dichotomous variables) between the matched groups. PSM appeared to have created remedial groups that were more similar on the covariate measures, reducing the preexisting difference between the groups however, groups were still unbalanced on some measures (see Tables 1 and 2).

A binary logistic regression analysis predicting remedial math-taking using the matched groups resulted in a model that provided a statistically significant prediction of remedial math-taking, $\chi^2(28, N = 3978) = 346.59$, $p <$ .001. The AUC was .68, which is considered poor discriminating ability, but better than tossing a coin (see Figure 1 for a comparison of the ROC curve before and after matching). Logistic regression resulted in six significant regression coefficients after matching compared to 10 before matching. The six significant predictors were all variables for which we found significant group differences had remained after matching. A comparison of logistic regression coefficients, standard errors, and odds ratios before and after matching are presented in Table 1. Results indicated that there were still differences in remedial math groups for some covariates after matching, and these differences significantly predicted remedial group membership. We did notice that the model Chi-square, Cox & Snell, Nagelkerke, predictive accuracy, and AUC values all decreased after matching, indicating a *worsened* model fit compared to the pre-matching model. The deterioration of the post-matching model, presumably, was the result of matched participants being more similar on the covariates, which was the goal of conducting PSM. However, PSM resulted in a 63% decrease in working sample size. It is possible that the results obtained may have been influenced by the missing data, particularly during PSM. We conducted a missing value analysis and multiple imputation to investigate the potential impact.

Table 1. Logistic regression summary: Covariates predicting remedial math before and after matching.

	Original Sample						12 Imputed Data Sets[b]					
	Before Matching (N = 7109)			After Matching (N = 3978)			Before Matching[a] (N = 10736)			After Matching (6356 ≤ N ≤ 6432)		
	B	S.E.	O.R.	B	S.E.	O.R.	B	S.E.	O.R.	B	S.E.	O.R.
CatholicHS	-.07	.09	.93	1.27	.14	3.55**	-.01	.08	.99	(1.16, 1.34)	.11	(3.17, 3.80)**
PrivateHS	-.37	.12	.69**	1.55	.21	4.69**	-.34	.09	.71**	(1.49, 1.67)	(.16, .17)	(4.44, 5.30)**
UrbanHS	.12	.10	1.12	.09	.11	1.09	.08	.08	1.08	(.07, .13)	.08	(1.08, 1.14)
Suburban	.07	.08	1.07	-.02	.09	.98	.10	.07	1.11	(-.01, .05)	.07	(.99, 1.06)
General	.20	.12	1.22	-.02	.12	.98	.21	.10	1.23*	(-.01, .12)	(.09, .10)	(.99, 1.13)
Coll.Prep	.12	.11	1.13	-.06	.12	.94	.16	.10	1.18	(-.03, .08)	.09	(.97, 1.08)
HiMath	-.70	.07	.50**	-.25	.08	.78**	-.70	.06	.50**	(-.24, -.21)	.06	(.79, .81)**
PSISector	-.91	.09	.40**	-.59	.10	.56**	-1.02	.07	.36**	(-.63, -.59)	.08	(.53, .55)**
Successful	-.34	.18	.71	-.17	.19	.85	-.28	.13	.76*	(-.22, -.08)	(.10, .14)	(.80, .92)
Marginal	.06	.17	1.06	-.11	.18	.90	.12	.12	1.13	(-.10, .00)	(.09, .13)	(.91, 1.00)
4-Year	-.63	.09	.53**	-.20	.10	.82*	-.59	.07	.56**	(-.28, -.23)	.08	(.76, .80)**
2-Year	-.41	.11	.67**	-.14	.12	.87	-.37	.09	.69**	(-.15, -.09)	(.09, .10)	(.87, .91)
2-Transfer	-.09	.11	.92	.03	.12	1.03	-.04	.09	.96	(-.02, .05)	.10	(.98, 1.05)
Black	-.08	.10	.93	.21	.11	1.23	.12	.08	1.13	(.22, .27)	.09	(1.25, 1.31)**
Hispanic	.31	.10	1.37**	.37	.12	1.44**	.27	.08	1.31**	(.23, .29)	.09	(1.25, 1.33)**
Asian	-.10	.13	.90	.14	.15	1.15	-.21	.10	.81*	(-.01, .04)	.11	(.99, 1.04)
Other	-.08	.14	.93	.01	.16	1.01	.06	.11	1.06	(.03, .10)	.12	(1.03, 1.11)
Single-Par.	.05	.08	1.05	.02	.08	1.02	.04	.06	1.04	(-.04, .01)	.07	(.97, 1.01)
Pipeline	-.49	.08	.61**	-.17	.09	.84	-.49	.07	.62**	(-.18, -.14)	.07	(.84, .87)*
Worked	.07	.07	1.07	-.03	.08	.97	.06	.06	1.06	(-.04, .00)	.06	(.96, 1.00)
Sex	.03	.06	1.03	.04	.07	1.04	-.02	.05	.98	(-.01, .02)	.05	(.99, 1.02)
Language	.16	.11	1.18	-.06	.12	.95	.09	.08	1.10	(-.11, -.07)	.09	(.89, .93)
Timing	.20	.09	1.22*	.07	.09	1.08	.21	.07	1.23**	(.05, .12)	.07	(1.05, 1.13)
Loans	.12	.07	1.13	.02	.07	1.02	.08	.06	1.09	(.02, .05)	.06	(1.02, 1.05)
Planning	.10	.07	1.10	.06	.08	1.06	.08	.06	1.08	(-.03, .04)	.06	(.97, 1.04)
SES	-.11	.05	.89*	-.01	.05	.99	.10	.04	1.11**	(-.04, -.01)	.04	(.97, .99)
Math	-.05	.00	.95**	-.01	.01	.99	.05	.00	1.05**	-.01	.00	.99**
Reading	.01	.01	1.01	.00	.01	1.00	-.01	.00	.99	.00	.00	1.00
Constant	1.72	.30	5.59	.59	.32	1.80	.91	.38	2.47	(.52, .72)	(.23, .24)	(1.69, 2.06)
Chi-Sq (df)	1853.55** (28)			346.59** (28)			(2634.55, 2664.32)** (28)			(538.40, 571.65)** (28)		
Hosmer-Lem.	26.76** (8)			54.00** (8)			(26.07, 37.06)** (8)			(66.27, 81.83)** (8)		
Cox & Snell	.23			.08			.22			(.08, .09)		
Nagelkerke	.32			.11			(.30, .31)			.11		
-2LL	6974.60			5168.09			(10904.29, 10934.07)			(8250.12, 8370.19)		
Accuracy	75.8%			63.5%			(74.4%, 74.7%)			(62.7%, 63.4%)		
AUC	.80			.68			.79			.68		

Notes. *$p < .05$, **$p < .01$. -2LL = -2 Log Likelihood. AUC = area under the ROC curve. a. Pooled. b. Values are reported as a range unless all 12 sets had an identical value.

Missing Value Analysis and Multiple Imputation

As missing values were present for multiple covariates, these cases were automatically deleted (listwise) during the preliminary analyses on group differences as well as during the PSM process. Fifteen variables had missing values and the ratio of the number of missing values to the number of cases was 2.25 to 1, and listwise deletion may have resulted in a substantial loss of information. Seven patterns of missing data occurred in more than 1% of the cases.

Little's MCAR test simultaneously compares differences between groups of missing and non-missing cases for every quantitative variable included and tests the null hypothesis that missing data are missing completely at

random. This test does not definitively determine the type of missing data, nor does it determine which variable(s) contributed to a significant result, but it does provide evidence that missing data may not be missing randomly (Meyers et al., 2013). Little's MCAR test was significant. MI was used to handle the missing values. All variables were included during the imputation process. The largest percentage of missing data was 12.10%, thus 12 imputations were performed.

Post-Imputation Analysis

We examined remedial group differences for the 15 variables with imputed data in all 12 MI sets. Test results showed significant differences between remedial math groups for all but one of the 15 variables and the resulting mean differences, standard errors, standardized residuals, and effect sizes were comparable to those found in the original working data set. Because there is no agreed upon method of pooling Chi-square statistics for multiply imputed data, we considered each statistic separately and reported the range of statistics and effect sizes in the 12 MI sets. This reporting procedure is used throughout the remainder of the analyses. See Tables 1 and 2 for a comparison of group differences before and after imputation.

The binary logistic regression with all covariates predicting remedial math-taking using the imputed data provided models that were significant predictors of remedial group in all 12 MI sets, $2634.55 \leq \chi^2(28,\ N = 10736) \leq 2664.32$, $p < .001$. The AUC was .79 (fair discriminating ability) across all imputed sets. See Table 1 for a summary of pooled logistic regression coefficients before and after MI. The histograms for the predicted probability of group membership for the MI sets had similar distributions to those from the original sample and are not presented. Overall, our post-imputation analyses found comparable remedial math group differences to those we found in the original sample. Imputing missing data did not have a significant effect on group differences. We next conducted PSM using the 12 MI sets.

Propensity Score Matching on All MI Sets

There were no missing values because we used the imputed data sets, thus the matching was conducted on a complete case sample in each MI set. Because one-to-one matching was used, the maximum number of matches that could be made was 3,509 (the lesser of the two frequencies of students in the remedial math groups). Just as we found in the original sample, remedial math groups were more similar after matching in the imputed sets. Additionally, a comparison of the histograms for the original and imputed data showed similar matching results thus they are not reported here. Overall, matching using the imputed data sets resulted in comparable results to those obtained by matching using the original sample. Imputing missing data did not appear to improve PSM results, although it did result in about a 62% increase in sample size compared to the matched sample in the original data.

Post-Matching Analyses of Remedial Group Differences

After matching using the MI sets, we found significant mean differences between remedial groups on all continuous variables as well as on 20 out of 30 nominal measures. A comparison of before and after matching revealed that differences and effect sizes were reduced. Once again, PSM appeared to have created remedial groups that were more similar on the covariate measures however, groups were still unbalanced (see Tables 1 and 2).

The binary logistic regression with all covariates predicting remedial math-taking using the imputed data provided models that were significant predictors of remedial group in all 12 sets, $538.40 \leq \chi^2(28) \leq 571.65$, $p < .001$.

The AUC was .68 for all models (less than acceptable discriminating ability) and this was identical to the result obtained using the matched groups in the original sample. Logistic regression resulted in nine significant regression coefficients after matching compared to 13 before matching. The nine significant predictors were all variables for which we found significant group differences had remained after matching. A comparison of logistic regression coefficients, standard errors, and odds ratios before and after matching are presented in Table 1.

Results indicated that there were still differences in remedial math groups for some covariates after matching using compete-case data, and these differences significantly predicted group membership. However, just as we saw in the original sample, the model Chi-square, Cox & Snell, Nagelkerke, predictive accuracy, and AUC values all decreased after matching, indicating a worsened model fit compared to the pre-matching model. Using PSM with the imputed data sets resulted in a 40% average decrease in working sample size compared to a 63% decrease when using the original data.

Remedial Math-Taking as the Sole Predictor

Binary logistic regression analysis with remedial math as the sole predictor produced a model that significantly predicted degree attainment in all four data samples (original, original-matched, imputed, and imputed-matched). Remedial math-taking was a significant negative predictor of degree attainment in all models. The predictive accuracy in the original data set was 65% compared to 57% in the original matched groups, and 56% in the

matched imputed samples. The AUC in the original data set was .61 (poor) compared to .56 (almost no discriminating ability) in all matched-group samples. The odds ratio for remedial math changed from .36 before matching to about .63 after matching. Although remedial math remained significant, the decrease in model accuracy, discriminating ability, and odds ratio imply that remedial math became less predictive of degree attainment after PSM.

Remedial Math with All Covariates as Predictors

Binary logistic regression using remedial math along with all covariates produced models that significantly predicted degree attainment in all four samples (original, original-matched, imputed, and imputed-matched). See Table 2 for a summary of the regression results for all covariates predicting degree attainment in all samples. There is no agreed upon method of pooling logistic regression statistics *across* models (Mood, 2010). Therefore, for the 12 matched imputed data sets, we considered each statistic separately and report the range of values if they were not identical in all 12 MI sets. Figure 4 shows side-by-side histograms comparing the predicted probability of degree attainment with all covariates included for all samples. For imputed samples, only the histogram for the first MI set is shown as those for the other 11 MI sets were similar.

Figure 3. Predicted probability of degree attainment including all covariates before and after matching and multiple imputation.

Before Matching After Matching

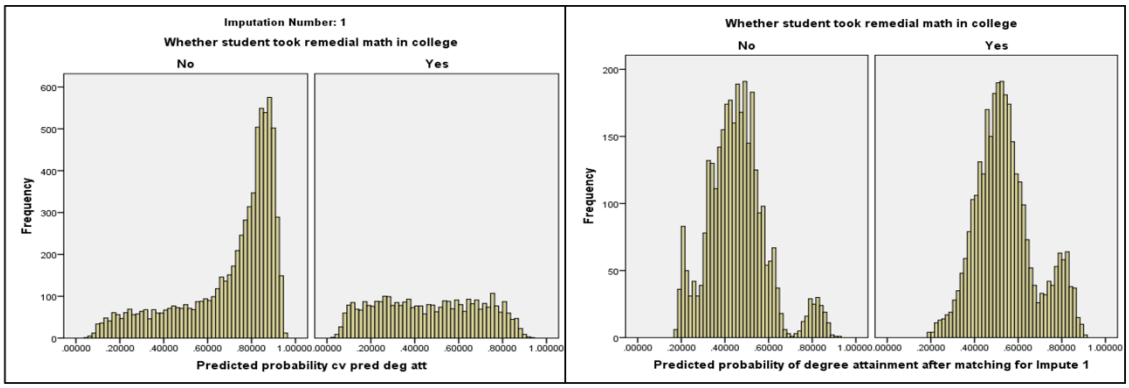


Table 2. Logistic regression summary: Covariates predicting degree attainment before and after matching

	Original Sample						12 Imputed Data Sets[b]					
	Before Matching (N = 7109)			After Matching (N = 3978)			Before Matching[a] (N = 10736)			After Matching (6356 ≤N≤ 6432)		
Predictor	B	S.E.	O.R.	B	S.E.	O.R.	B	S.E.	O.R.	B	S.E.	O.R.[c]
Remedial	-.38	.07	.68**	-.41	.08	.67**	-.36	.05	.70**	(-.36, -.32)	.06	(.7, .73)**
CatholicHS	.26	.09	1.29**	.17	.14	1.19	.19	.08	1.21**	(.04, .11)	.11	(1.04, 1.11)
PrivateHS	.27	.11	1.31*	.37	.19	1.44*	.14	.09	1.15	(.14, .21)	(.14, .15)	(1.15, 1.24)
UrbanHS	-.09	.10	.92	-.07	.12	.94	.01	.08	1.01	(-.05, .05)	.09	(.95, 1.05)
Suburban	.03	.08	1.03	.01	.10	1.01	.06	.07	1.06	(.02, .08)	.08	(1.02, 1.09)
General	-.03	.12	.97	.07	.14	1.07	-.02	.10	.98	(-.08, .09)	.1	(.92, 1.09)
Coll Prep	.00	.12	1.00	.06	.13	1.07	.01	.10	1.01	(-.06, .14)	.10	(.94, 1.15)
HiMath	.41	.07	1.51**	.46	.08	1.58**	.34	.06	1.41**	(.28, .36)	.07	(1.33, 1.43)**
PSISector	.19	.08	1.21*	.13	.10	1.14	.25	.06	1.29**	(.14, .21)	.08	(1.16, 1.23) [8*]
Successful	1.34	.22	3.83**	1.19	.24	3.27**	.94	.21	2.55**	(.66, 1.19)	(.11, .17)	(1.93, 3.27)**
Marginal	.70	.21	2.00**	.64	.23	1.89**	.33	.21	1.39	(.08, .58)	(.1, .16)	(1.09, 1.79) [11*]
4-Year	.42	.09	1.52**	.47	.11	1.59**	.34	.07	1.40**	(.31, .37)	.09	(1.36, 1.45)**
2-Year	-.39	.11	.68**	-.30	.13	.74*	-.44	.08	.64**	(-.4, -.35)	.1	(.67, .71)**
2-Transfer	.78	.11	2.17**	.91	.13	2.49**	.76	.09	2.15**	(.77, .88)	.11	(2.16, 2.41)**
Black	-.38	.10	.68**	-.34	.12	.71**	-.47	.08	.63**	(-.54, -.44)	(.09, .1)	(.59, .64)**
Hispanic	-.11	.11	.89	-.07	.13	.94	-.23	.08	.80**	(-.28, -.15)	.10	(.76, .86) [10*]
Asian	.24	.13	1.27	.23	.17	1.26	.08	.10	1.08	(-.06, .03)	.12	(.94, 1.03)*
Other	-.42	.13	.66**	-.34	.17	.71*	-.35	.11	.70**	(-.44, -.32)	.13	(.64, .73)*
Single-Par.	-.07	.08	.94	-.13	.09	.88	-.08	.06	.93	(-.13, -.07)	.07	(.88, .93)
Pipeline	.38	.08	1.46**	.45	.10	1.57**	.33	.07	1.40**	(.3, .41)	.07	(1.35, 1.51)**
Worked	.05	.07	1.05	.15	.09	1.16	-.03	.06	.97	(.03, .1)	.07	(1.03, 1.1)
Sex	.45	.06	1.56**	.37	.08	1.45**	.45	.05	1.56**	(.37, .41)	.06	(1.44, 1.51)**
Language	-.11	.11	.90	-.10	.14	.91	-.16	.08	.85	(-.2, -.13)	.10	(.82, .88) [2*]
Timing	.79	.09	2.21**	.73	.10	2.07**	.80	.07	2.22**	(.72, .83)	.07	(2.05, 2.29)**
Loans	.50	.06	1.65**	.53	.08	1.71	.47	.05	1.60**	(.48, .56)	.06	(1.62, 1.76)**
Planning	-.06	.07	.94	-.15	.09	.86	-.03	.06	.97	(-.07, .01)	.07	(.93, 1.01)
SES	.27	.05	1.31**	.22	.06	1.25**	.27	.04	1.32**	(.18, .24)	.05	(1.2, 1.28)**
Math	.01	.00	1.01	.01	.01	1.01	.01	.00	1.01*	(0, .01)	.00	(1, 1.01) [3*]
Reading	-.02	.01	.98**	-.02	.01	.99**	-.02	.00	.98**	(-.02, -.01)	.01	(.98, .99)**
Constant	-1.67	.33	.19	-1.70	.38	.18	-1.32	.29	.27	(-1.56, -1.07)	(.25, .28)	(.21, .34)
Chi-Sq(df)	1968.27 (29)**			1033.15** (29)			(2975.57, 3017.08)** (29)			(1504.39, 1585.39)** (29)		
Hosmer & Lem.	7.94 (8)			11.42 (8)			(16.03, 23.83)* (8)			(1.61, 2.88) (8) [4*]		
Cox & Snell	.24			.23			(.24, .25)			(.21, .22)		
Nagelkerke	.34			.31			.33			(.28, .29)		
-2LL	7068.23			4401.87			(11178.78, 1122.29)			(7214.72, 7348.74)		
Accuracy	76.7%			72.6%			(75.0%, 75.3%)			(71.1%, 72.1%)		
AUC	.80			.79			.80			(.76, .78)		

Notes. *$p < .05$, **$p < .01$. -2LL = -2 Log Likelihood. AUC = area under the ROC curve. a. Pooled. b. Values are reported as a range unless all 12 sets had an identical value. c. If not sig. in all 12 imputed matched sets, [ ] indicates number of sets in which the predictor/statistic was sig. at $\alpha \leq .05$

See Figure 1 for a comparison of the ROC curves for regression analyses before and after matching. The AUC in the original and imputed samples was .80 (good), compared to .79 (fair) in the original matched, and .77 (fair) in the matched-imputed, samples. Remedial math, when combined with all covariates, was a significant negative predictor of degree attainment in all four full models. Overall, the predictive accuracy and discriminating ability of remedial math-taking combined with all covariates remained relatively consistent in all models, as did the unique negative effect of remedial math-taking. Imputing the missing values and matching participants resulted in very little change in the estimates of treatment effects on degree attainment.

Discussion

The purpose of this study was to investigate whether multiple imputation procedures and propensity score matching would improve estimates of the treatment effects of remedial math-taking on degree attainment. PSM is purportedly a superior method for creating equivalent comparison groups in non-experimental studies. In this study, matching seemed to help make groups of participants more similar on covariate measures. We were able to reduce the magnitude of group differences, but still had difficulty achieving balance between matched groups despite the inclusion of numerous covariates. One possible explanation for our findings of significant group differences post-matching was our large sample sizes. Standard procedures for investigations of group differences (*t*-test, Chi-square) often produce significant results given a large enough sample size. An examination of the effect sizes for significant results revealed relatively small effect sizes for group differences after matching. However, given the uncertainty as to what constitutes negligible group imbalance (Austin, 2009), evaluating the effectiveness of PSM is subjective. Several noticeable differences in models pre- and post-matching were found. The discriminating ability of the regression models that used the covariate measures to predict group membership worsened after matching, indicating that group membership was less influenced by the covariate measures. However, remedial math alone was a poor predictor of degree attainment (as measured by the AUC) regardless of whether the samples were matched, or missing data were imputed. With the inclusion of all covariates predicting degree attainment, the full models had relatively similar discriminating ability, regardless of matching or imputation. Although remedial math was consistently negatively associated with degree attainment in all models, the unique contribution of remedial math, as measured by its odds ratio, was significantly reduced after matching. However, these results did not allow us to definitively say that this change was the result of PSM doing its job.

Ultimately, PSM did not appear to improve the estimates of treatment effects over and above those we obtained using the standard logistic regression analyses conducted on the unmatched sample using all covariates as predictors. Our results were similar to those found by other authors (Byun et al., 2015; Melguizo et al., 2011, Padgett et al., 2010). Moreover, matching resulted in a significant decrease in the working sample size, while producing no substantial improvement in effect estimates. Although missing value analyses indicated that missing data might have produced biased estimates of treatment effects, regression analyses pre- and post-matching on the multiply imputed data sets showed similar results to the previous analyses on the original data set.

Running a single round of 12 imputations took approximately one and a half hours each time. In addition, systematically adding and removing covariate measures to the PSM analysis in an attempt to achieve balance between matched groups took several weeks. Given that these methods did not provide significantly different estimates of treatment effects, the length of time it took to conduct the analyses is a deterrent to using them again. Furthermore, statisticians disagree on an adequate way to pool results of analyses on multiply imputed data sets. Conducting PSM on the MI data required us to conduct the logistic regression analysis individually on each matched group (12 groups) and the results had to be evaluated separately. PSM and MI themselves were limiting factors in this study. As PSM and MI rely heavily on selection of adequate covariates, it is plausible that not all relevant covariate measures were included in the analysis, but there is no way of knowing in advance whether adequate covariates were selected. We used a one-to-one matching technique and recommend future studies investigate alternative matching methods.

References

Attewell, P., Lavin, D., Domina, T., & Levey, T. (2006). New evidence on college remediation. *Journal of Higher Education*, *77*, 886–924.

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*(25), 3083–3107.

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

Bahr, P. R. (2007). Double jeopardy: Testing the effects of multiple basic skill deficiencies on successful remediation. *Research in Higher Education*, *48*, 695–725.

Bahr, P. R. (2010). Preparing the underprepared: An analysis of racial disparities in postsecondary mathematics remediation. *The Journal of Higher Education*, *81*, 209–237.

Bettinger, E., & Long, B. T. (2005). Remediation at the community college: Student participation and outcomes. *New Directions for Community Colleges*, *2005*(129), 17–26.

Bonham, B. S., & Boylan, H. R. (2012). Developmental mathematics: Challenges, promising practices, and recent initiatives. *Journal of Developmental Education*, *36*(2), 14–21.

Bozick, R., Lauff, E., & Wirt, J. (2007). Education longitudinal study of 2002 (ELS: 2002): A first look at the initial postsecondary experiences of the high school sophomore class of 2002. *National Center for Education Statistics*.

Byun, S. Y., Irvin, M. J., & Bell, B. A. (2015). Advanced math course taking: Effects on math achievement and college enrollment. *The Journal of Experimental Education*, *83*, 439–468.

Carlin, J. B., Li, N., Greenwood, P., & Coffey, C. (2003). Tools for analyzing multiple imputed datasets. *The Stata Journal*, *3*(3), 226–244.

Clark, M., & Cundiff, N. (2011). Assessing the effectiveness of a college freshman seminar using propensity score adjustments. *Research in Higher Education*, *52*, 616–639.

Fan, X., & Nowell, D. L. (2011). Using propensity score matching in educational research. *Gifted Child Quarterly*, *55*, 74–79.

Giani, M., Alexander, C., & Reyes, P. (2014). Exploring variation in the impact of dual-credit coursework on postsecondary outcomes: A quasi-experimental analysis of Texas students. *The High School Journal*, *97*(4), 200–218.

Harel, O. (2007). Inferences on missing information under multiple imputation and two-stage multiple imputation. *Statistical Methodology*, *4*(1), 75–89.

Harel, O., & Zhou, X. H. (2007). Multiple imputation: Review of theory, implementation and software. *Statistics in Medicine*, *26*(16), 3057–3077.

Henderson, J., & Chatfield, S. (2011). Who matches? Propensity scores and bias in the causal effects of education on participation. *Journal of Politics*, *73*, 646–658.

Hill, J. (2004). *Reducing bias in treatment effect estimation in observational studies suffering from missing data* (Working Paper 04–01). Columbia University Academic Commons.

Hill, J., Weiss, C., & Zhai, F. (2011). Challenges with propensity score strategies in a high-dimensional setting and a potential alternative. *Multivariate Behavioral Research*, *46*(3), 477–513.

Holmes, W. M. (2014). *Using propensity scores in quasi-experimental designs*. Los Angeles, CA: SAGE Publications.

Illich, P. A., Hagan, C., & McCallister, L. (2004). Performance in college-level courses among students concurrently enrolled in remedial courses: Policy implications. *Community College Journal of Research & Practice, 28*, 435–453.

Luellen, J. K., Shadish, W. R., & Clark, M. H. (2005). Propensity scores: An introduction and experimental test. *Evaluation Review, 29*, 530–558.

Melguizo, T., Kienzl, G. S., & Alfonso, M. (2011). Comparing the educational attainment of community college transfer students and four-year college rising juniors using propensity score matching methods. *Journal of Higher Education, 82*, 265–291.

Meyers, L. S., Gamst, G., & Guarino, A. J. (2013). *Applied multivariate research: Design and interpretation* (2nd ed.). Los Angeles, CA: SAGE Publications.

Miles, A. (2016). Obtaining predictions from models fit to multiply imputed data. *Sociological Methods & Research, 45*(1), 175-185.

Mitra, R., & Reiter, J. P. (2016). A comparison of two methods of estimating propensity scores after multiple imputation. *Statistical methods in medical research, 25*(1), 188–204.

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review, 26*(1), 67–82.

Normand, S. L. T., Landrum, M. B., Guadagnoli, E., Ayanian, J. Z., Ryan, T. J., Cleary, P. D., & McNeil, B. J. (2001). Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. *Journal of Clinical Epidemiology, 54*(4), 387–398.

Padgett, R. D., Salisbury, M. H., An, B. P., & Pascarella, E. T. (2010). Required, practical, or unnecessary? An examination and demonstration of propensity score matching using longitudinal secondary data. *New Directions for Institutional Research, 2010*, 29–42.

Peikes, D. N., Moreno, L., & Orzol, S. M. (2008). Propensity score matching: A note of caution for evaluators of social programs. *The American Statistician, 62*(3), 222–231.

Pretlow III, J., & Washington, H.D. (2011), Cost of developmental education: An update to Breneman and Harlow. *Journal of Developmental Education, 35*(1), 2–12.

Reist, B. M., & Larsen, M. D. (2012). Post-Imputation Calibration Under Rubin's Multiple Imputation Variance Estimator. In *Section on Survey Research Methods, Joint Statistical Meeting*.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*(1), 33–38.

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

Titus, M. A. (2007). Detecting selection bias, using propensity score matching, and estimating treatment effects: An application to the private returns to a master's degree. *Research in Higher Education, 48*(4), 487–521.

Vaughan, A., Lalonde, T., & Jenkins-Guarnieri, M. (2014). Assessing student achievement in large-scale educational programs using hierarchical propensity scores. *Research in Higher Education, 55*, 564–580.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine, 30*(4), 377–399.

Yuan, Y. C. (2010). Multiple imputation for missing data: Concepts and new development (Version 9.0). *SAS Institute Inc., Rockville, MD, 49*, 1–11.