ASSESSMENT

Impact of Undergraduate Research Training Programs: An Illustrative Example of Finding a Comparison Group and Evaluating Academic and Graduate School Outcomes

Kaitlyn N. Stormes, *University of California, Los Angeles*
Nicole A. Streicker, *California State University, Long Beach*
Graham K. Bowers, *University of California, Los Angeles*
Perla Ayala, Guido G. Urizar Jr., *California State University, Long Beach*

Abstract

In this study, researchers at a large, urban, comprehensive minority-serving institution used propensity score matching to identify a unique comparison group to study academic and graduate school outcomes in students served by the National Institutes of Health–funded Building Infrastructure Leading to Diversity (BUILD) Initiative. Acknowledging that students' self-selection biases may confound findings, the use of propensity methods to match students served with those who were not (but were otherwise eligible) provides a valuable tool for evaluators and practitioners to combat this challenge and better evaluate their effectiveness and impact on students' success. This study's findings indicate that BUILD participants had higher academic and graduate school success with regard to cumulative GPA, units attempted and completed, graduation status, and application and admission to graduate programs.

Keywords: *academic success; graduate school outcomes; NIH–funded undergraduate training programs; propensity score matching; undergraduate research training programs*

doi: 10.18833/spur/5/3/8

First-generation students, racial and ethnic minority students, students with disabilities, and those from other disadvantaged backgrounds are severely underrepresented in biomedical sciences, engineering, and behavioral health sciences (BSE/BHS; NSB 2012; Nelson 2004). National US data show that these underrepresented groups (URGs) represent only 14 percent of earned baccalaureate degrees in science and 6 percent of BSE/BHS doctorates, compared with 81 percent and 74 percent for their non-URG peers (NCES 2001; NSB 2012). To address this lack of diversity, the inclusion of URGs in undergraduate research-training programs (URTPs) is recognized as a significant step toward engaging and retaining URGs in health-related research careers.

Engaging in research has been associated with improvements in student learning and critical thinking (Brownell et al. 2015), which in turn have led to increases in cumulative GPA (Haeger and Fresquez 2016), retention and persistence in undergraduate science, technology, engineering, and math (STEM) degrees (Russell, Hancock, and McCullough 2007), shorter time-to-degree (Kinkel and Henke 2006), and higher graduation rates (Jones, Barlow, and Villarejo 2010). Improvements in these academic outcomes are even greater among URGs the longer they engage in URTPs (Hernandez et al. 2018). URTPs also strengthen students' preparation for BSE/BHS graduate studies, increasing their probability of pursuing and enrolling in a BSE/BHS graduate program and generating valued products for graduate schools, including publications, presentations, and awards (Eagan et al. 2013; Weston and Laursen 2015; Wilson et al. 2018).

Despite these promising findings, quantitative assessments of URTPs are rare due to the challenging process of tracking long-term student outcomes (Linn et al. 2015). Prior literature suggests that between 29 percent (Junge et al. 2010) and 33 percent (Hall 2017) of students in formal URTPs complete PhD degrees. However, few studies report the

impact of structured URTPs with a matched comparison group on graduate school application, admission, and matriculation outcomes. Furthermore, few studies document the nuances of these outcomes by comparing findings on URGs and non-URGs. Including a comparison group when testing the effectiveness of URTPs has been limited due to challenges in implementing traditional randomized controlled trials in higher education, with randomization of students to support programs creating ethical and practical dilemmas for evaluators and researchers (Cook 2001).

Therefore, the development of improved and dependable evaluation methods to assess program effectiveness has become a priority. As states move from enrollment-based to performance-based funding models, institutions and programs that effectively meet their accountability measures will be prioritized (Boggs 2018; Dougherty et al. 2014; LAO 2007). Relying on descriptive statistics or only reporting numbers of students supported by these programs who meet the intended goals misses the opportunity for comprehensive evaluation (Stuart and Rubin 2008; Weston and Laursen 2015). Higher education researchers and evaluators must use techniques to develop adequate comparison groups and rule out selection bias to demonstrate program success, secure funding, and disseminate more robust findings.

One solution to developing adequate comparison groups is to use statistical approaches that match program participant characteristics to similar students who did not participate in the program but were otherwise eligible. Propensity score matching (PSM) provides an opportunity to match students based on a propensity score: "the conditional probability of assignment to a particular treatment given a vector of observed covariates" (Rosenbaum and Rubin 1985). Using this approach, researchers can rule out selection bias and more accurately assess differences in program outcomes of interest, such as student retention and graduate school application, acceptance, and matriculation (Angrist and Pischke 2009). Propensity score matching is a statistical estimation of group enrollment and can be used to select cases that are statistically similar to an intervention group. Rather than predicting who may or may not be enrolled, it selects a subsample of cases that include a covariate factor to reduce the strength of the effect of confounding variables in final analyses by creating a more similar sample for comparison.

Despite these benefits, there are some requirements to PSM that may explain why few published studies have reported using propensity methods to evaluate URTPs that support underserved students. To find comparison group matches with adequately matching scores, PSM requires large datasets with no missing data; URTP research often includes missing data and smaller sample sizes. There also may be a lack of knowledge regarding its utility due to its slow migration of use from health sciences to social and educational research. However, with complete data and increased sample sizes, propensity methods allow the creation of pseudo-randomized control groups that temper the effects of possible confounds, enabling meaningful analyses and results.

The purpose of this study is to provide an illustrative example of using PSM to identify a comparison group and test the effectiveness of a URTP (California State University, Long Beach Building Infrastructure Leading to Diversity Initiative, or CSULB BUILD) in enhancing student academic and graduate school outcomes. The following hypotheses guide this study:

1. BUILD students, when compared to matched non-BUILD students, will have higher cumulative grade point averages (GPAs), more units attempted and units earned, will be more likely to graduate, and have less time-to-degree.
2. BUILD students, when compared to matched non-BUILD students, will apply to, get accepted to, and attend graduate and professional programs at higher rates. The authors also explored whether these outcomes differed by URG status and PhD programs specifically.

Method

Data Collection

This study analyzed existing program and institutional data (e.g., demographic and academic outcomes) and additional primary data (i.e., graduate school outcomes) on a sample of undergraduate students who matriculated at California State University, Long Beach (CSULB) between fall 2010 and spring 2019. Institutional Review Board (IRB) approval was obtained before data collection and analysis.

Demographic and Academic Data

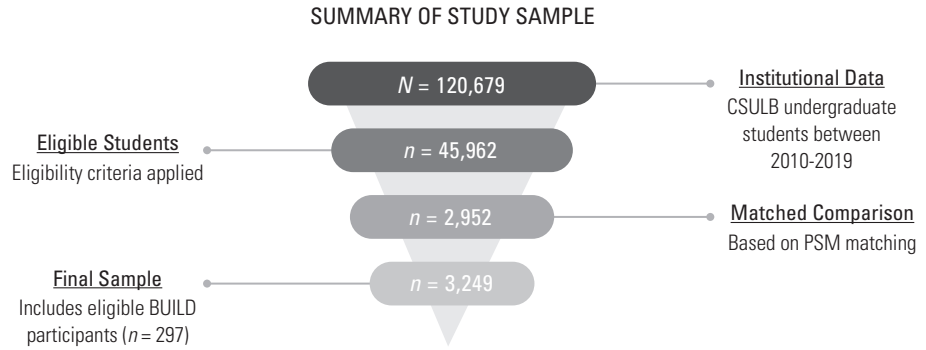
The local institutional research office provided student demographic information and admission, enrollment, and degree attainment data. BUILD participant programmatic and graduate school outcome data were acquired directly from the students, from their mentors, and by other data collection efforts (i.e., BUILD Alumni Survey, follow-up emails and outreach, and additional annual report records).

To determine if non-BUILD students participated in undergraduate research, data were provided from academic colleges, the local institutional research office, and the Office of Research and Economic Development (i.e., college safety-training rosters, research program rosters, enrollment in inquiry-based research courses, and Collaborative Institutional Training Initiative records).

Graduate School Data

Graduate school outcome data was collected directly from CSULB URTPs and through an online Students

FIGURE 1. Participant flowchart



and Alumni Survey that was administered to BUILD and non-BUILD students (n = 3,214). One invitation and weekly reminders were sent, with the survey remaining open for four weeks. Consenting participants answered 11 questions, including, "Were you admitted to any of the programs to which you applied?" and "If you applied and were admitted, did you attend, or do you plan to attend?" Of the 3,214 students invited to complete the survey, 636 (20 percent) participated.

Matched Comparison Group

To identify a BUILD comparison group, a 1:10 nearest neighbor matching without replacement PSM method was used. Although optimal matching typically results in more matched cases, nearest neighbor matching is considered more robust (Rosenbaum and Rubin 1985; Stuart and Rubin 2008) and allows students to be matched on propensity scores at a stricter threshold (Caliendo and Kopeinig 2008). A caliper (the maximum permitted statistical difference between matches) of 0.20 was selected (Rosenbaum and Rubin 1985; Stuart and Rubin 2008). Non-BUILD students who did not meet all eligibility requirements were excluded from consideration for a match. Of 120,769 undergraduate students, 45,962 students met BUILD initiative eligibility criteria and were used for initial PSM (see Figure 1).

Next, logistic regression models were used to predict the probability of students participating in BUILD, and covariates were selected based on the literature (Caliendo and Kopeinig 2008; Lane et al. 2012). All covariates with the exception of gender, Pell grant eligibility, continuous enrollment, and some cohorts and colleges were significant (see Table 1).

Participants

The PSM analysis resulted in a 1:10 match of 297 BUILD to 2,952 non-BUILD students. After propensity scores were estimated and participants were matched, standardized mean differences were analyzed to assess the match quality of BUILD and non-BUILD students (Caliendo and Kopeinig 2008). After using PSM, BUILD and non-BUILD students were more similar, with the standard mean difference for student characteristics, such as full-time enrollment (i.e., 12 or more units), dropping from 0.989 to 0.012 (see Table 2). This increase in match between groups can be seen across variables, illustrating that all covariates were now more balanced, with all standard mean differences less than 0.1 after matching (see Table 3 for full demographics of BUILD and non-BUILD students).

Program Description

At CSULB, health-related research is interdisciplinary and includes basic, applied, and translational approaches to studying a variety of prominent health issues. CSULB BUILD has pursued a range of strategies to encourage preparation and persistence in health-related disciplines, including providing a culturally sensitive student training program, rigorous health-related research training, mentorship opportunities for URGs (Abeywardana et al. 2020), and training for faculty mentors (Urizar et al. 2017; Young and Stormes 2020).

Analyses

A priori power analyses were conducted using G-Power to ensure that minimum sample sizes were obtained and to avoid a type II error. Adequate power of 0.80 was present for all analyses and sample sizes proposed. For the mean comparison independent samples t-tests, with an alpha of 0.05, power of 0.95, and Cohen's d of 0.60, a minimum sample of 122 per group was needed. A Cohen's d of 0.20, 0.50, and 0.80 represent a small, medium, and large effect, respectively (Cohen 1988). For the chi-square nonparametric tests, with an alpha of 0.05 and power of 0.80, and ϕ of 0.5, the minimum sample size needed was 32 per group. Cramer's ϕ ranges from zero to one where 0.10, 0.30, and 0.50 represent a small, medium, and large effect, respectively. This study exceeded these recommended minimum sample size requirements.

TABLE 1. Logistic Regression Predicting BUILD Participation (*n* = 45,962)

Variables	B	SE	*p* value	Exp(B)
Gender (female)	-0.131	0.120	.277	0.878
Full-time enrollment (12 units)	-2.119	0.359	.000***	0.120
Academic cohort year				
2010–2011	-1.520	0.791	.055	0.219
2011–2012	-2.136	1.061	.044*	0.118
2012–2013	0.258	0.457	.572	1.295
2013–2014	1.056	0.401	.008**	2.876
2014–2015	2.053	0.371	.000***	7.789
2015–2016	1.908	0.375	.000***	6.531
2016–2017	1.876	0.377	.000***	6.531
2017–2018	1.668	0.388	.000***	5.300
Number of terms enrolled	0.085	0.020	.000***	1.088
Age at entry	-0.144	0.024	.000***	0.866
Student type (first-time first-year students)	1.171	0.138	.000***	3.224
Academic college				
Business	-15.298	1231.039	.990	0.000
Education	-15.298	2734.785	.996	0.000
Engineering	1.363	1.009	.996	0.000
Health & Human Services	0.273	1.011	.177	3.908
Liberal Arts	0.394	1.008	.787	1.314
Natural Sciences & Mathematics	2.263	1.007	.025*	9.611
Arts	-0.019	1.416	.989	0.981
Pell grant eligibility	0.072	0.119	.546	1.075
Pell grant amount (dollars)	0.000	0.000	.004**	1.000
Minority status	-0.532	0.134	.000***	0.588
Math remediation needed	0.783	0.284	.006**	2.188
English remediation needed	0.732	0.257	.004**	2.079
First-generation status	0.261	0.117	.025*	1.298
Continuously enrolled	-0.071	0.251	.777	0.932
Change of academic college	-4.996	0.065	.000***	0.007

Note: Academic year 2018–2019 is the reference group for the cohort and undeclared is the reference group for academic college. B = regression coefficient, SE = Standard error, and Exp(B) = Odds ratio.
*$p < .05$, **$p < .01$, ***$p < .001$.

Independent samples *t*-tests and Pearson's chi-square tests were used to analyze between-group differences of continuous and categorical dependent variables, respectively. For both types of analyses, BUILD student status (BUILD or non-BUILD) was the independent variable, and academic (i.e., units attempted, units earned, units completed, time-to-degree, and college change), and graduate school outcomes (i.e., application, admission, and matriculation) were the dependent variables. Additional post hoc analyses examined the extent to which BUILD equitably supported URGs across the same outcomes.

Results

Academic Outcomes

BUILD students attempted more units ($M = 88.92$, $SD = 33.14$) and earned more units ($M = 86.20$, $SD = 32.64$) than non-BUILD students ($M = 83.31$, $SD = 39.02$; $M = 79.97$, $SD = 37.78$). However, BUILD and non-BUILD students did not differ in completion rate (i.e., the percent of units earned out of units attempted). Most notably, BUILD students had higher cumulative GPAs ($M = 3.42$, $SD = 0.39$) than non-BUILD students ($M = 3.17$,

TABLE 2. Matching Covariate Balance before and after Propensity Score Matching

	Unmatched			Matched		
Variables	BUILD	Non-BUILD	SMD	BUILD	Non-BUILD	SMD
Size (*n*)	297	2,952	–	297	2,952	–
Female students	0.623 (0.485)	0.592 (0.492)	0.064	0.623 (0.485)	0.613 (0.487)	0.021
Full-time enrollment (12 units)	0.973 (0.162)	0.813 (0.39)	0.989	0.973 (0.162)	0.971 (0.168)	0.012
Academic year cohort	1508.859 (150.532)	1405.954 (253.189)	0.684	1508.859 (150.532)	1511.118 (190.789)	0.015
Number of terms enrolled at CSULB	6.296 (2.342)	5.575 (2.89)	0.308	6.296 (2.342)	6.177 (2.948)	0.051
Age at entry	19.707 (3.321)	21.255 (4.579)	0.466	19.707 (3.321)	19.775 (3.66)	0.02
Type of student	1.229 (0.421)	1.489 (0.5)	0.618	1.229 (0.421)	1.236 (0.425)	0.018
College	4.721 (1.177)	4.429 (1.069)	0.247	4.721 (1.177)	4.736 (1.131)	0.013
Pell grant eligibility	0.606 (0.489)	0.623 (0.485)	0.035	0.606 (0.489)	0.593 (0.491)	0.027
Pell amount offered in dollars	9072.333 (9622.262)	7609.026 (8581.653)	0.152	9072.333 (9622.262)	8629.545 (9689.465)	0.046
Minority status	0.599 (0.491)	0.527 (0.499)	0.148	0.599 (0.491)	0.587 (0.492)	0.024
Math remediation needed	0.044 (0.205)	0.091 (0.288)	0.231	0.044 (0.205)	0.038 (0.192)	0.026
English remediation needed	0.054 (0.226)	0.106 (0.308)	0.23	0.054 (0.226)	0.05 (0.218)	0.018
First-generation student	0.478 (0.5)	0.543 (0.498)	0.13	0.478 (0.5)	0.465 (0.499)	0.027
Continuously enrolled	0.943 (0.233)	0.939 (0.24)	0.017	0.943 (0.233)	0.945 (0.228)	0.01
Change of college	0.205 (0.405)	0.236 (0.424)	0.075	0.205 (0.405)	0.203 (0.402)	0.006

Note: Good covariate balance = standardized mean difference (SMD) of < 0.1.

SD = 0.43), $t(3,247) = 9.447$, $p < .001$, Cohen's $d = 0.61$ (see Table 4 for all *t*-test results). The magnitude of this effect was medium (Cohen 1988), demonstrating a considerable difference between BUILD and non-BUILD cumulative GPAs.

Additionally, BUILD students were significantly more likely to graduate from college compared to non-BUILD students (52 percent and 33 percent, respectively), $\chi^2(1) = 43.50$, $p < .001$, $\phi = 0.116$. However, among those who graduated, there were no group differences in the frequency of changing degree or in time-to-degree (measured in number of semesters).

Further analyses were conducted to examine whether the BUILD Initiative had a differential impact on URGs. Differences between BUILD and non-BUILD students in units attempted and earned were no longer significant when accounting for URG status (see Table 4). However, GPA differences remained significant such that both BUILD URG ($M = 3.37$, $SD = 0.36$) and non-URG students ($M = 3.54$, $SD = 0.35$) had higher cumulative GPAs than non-BUILD URG [$M = 3.19$, $SD = 0.39$; $t(642) = 4.09$, $p < .001$, Cohen's $d = 0.48$] and non-URG counterparts [$M = 3.29$, $SD = 0.38$; $t(485) = 4.85$, $p < .001$, Cohen's $d = 0.68$; see Table 4].

Regarding graduation, both BUILD URG ($\chi^2(1) = 28.30$, $p < 0.001$, $\phi = 0.122$) and non-URG students ($\chi^2(1) = 15.55$, $p < .001$, $\phi = 0.108$) graduated at higher rates (52 percent and 53 percent, respectively) than non-BUILD URG and non-URG students (32 percent and 35 percent, respectively). Additionally, there were no significant between-group differences by URG degree earners in the frequency of changing degree or in time-to-degree.

Graduate School Outcomes

Because graduate school outcomes are an important mission of URTPs, BUILD and non-BUILD students were compared on application, admission, and matriculation to graduate school in general (master's and PhD programs), and specifically doctoral programs. These analyses were limited to those who graduated or who were graduate-school eligible, narrowing the available sample for comparison to 155 BUILD students and 257 non-BUILD students.

TABLE 3. Demographic Composition of BUILD and Non-BUILD Matched Samples

Description	BUILD	Non-BUILD
Size (*n*)	297	2,952
Age	*M (SD)*	*M (SD)*
	19.71 (3.321)	19.78 (3.669)
Gender	*n* (%)	*n* (%)
Male	112 (37.7%)	1,146 (38.8%)
Female	185 (62.3%)	1,806 (61.2%)
Pell grant eligibility	180 (60.6%)	1,749 (59.2%)
First-generation student	142 (47.8%)	1,373 (46.5%)
Underrepresented minority	178 (59.9%)	1,731 (58.6%)
College		
Natural Sciences & Mathematics	98 (33.0%)	554 (18.8%)
Liberal Arts	79 (26.6%)	1,130 (38.3%)
Engineering	68 (22.9%)	421 (14.3%)
Health & Human Services	50 (16.8%)	714 (24.2%)
Arts	1 (0.3%)	54 (1.8%)
University programs (Undeclared)	1 (0.3%)	58 (2.0%)
Education	0 (0%)	4 (0.1%)
Business	0 (0%)	17 (0.6%)

Note: There were no significant differences between BUILD and Non-BUILD students. First-generation students are defined as "undergraduates whose parents never enrolled in postsecondary education" (Cataldi, Bennett, and Chen 2018). Pell grant eligibility was used as a proxy for low-income status (Rosinger and Ford 2019). Underrepresented minority was defined as belonging to one of the following race/ethnicity categories: Black and African American, Hispanic or Latino, American Indian or Alaska Native, Native Hawaiian, Other Pacific Islander, Cambodian, Hmong, or Laotian (NSF 2019; Teranishi, Lok, and Nguyen 2013).

BUILD students were more likely to apply to any graduate program (75 percent) than non-BUILD students [42 percent; $\chi^2(1) = 48.51$, $p < .001$, $\phi = 0.32$]. Among those that applied, BUILD students ($n = 103$; 79 percent) were more likely to be admitted to graduate school than non-BUILD students [$n = 77$; 61 percent; $\chi^2(1) = 10.59$, $p < .001$, $\phi = 0.20$]. However, among those admitted, there were no group differences in graduate program matriculation. For doctoral graduate programs, more BUILD students (97 percent) applied than non-BUILD students [68 percent; $\chi^2(1) = 24.8$, $p < .001$, $\phi = 0.42$]. However, there was no difference between BUILD and non-BUILD PhD program admission or matriculation.

When examining graduate school outcomes by URG status, BUILD URG and non-URG students had higher application (79 percent and 70 percent, respectively) and admission (77 percent and 83 percent, respectively) to graduate programs compared to non-BUILD URG and non-URG students (application: 47 percent and 36 percent, respectively; admission: 58 percent and 65 percent, respectively), but did not differ in matriculation. Findings were similar regarding doctoral program application, with BUILD students, regardless of URG status, having applied at higher rates (97 percent) than non-BUILD students (URG: 65 percent; non-URG: 71 percent; see Table 5 for chi-square results).

Discussion

Academic Outcomes

Results demonstrated that although BUILD students attempted and earned more units than non-BUILD students, they had comparable completion rates. These patterns were observed for both URG and non-URG students and are similar to previous studies showing that participation in URTPs (particularly for one academic year or longer) led to increased persistence, improvements in cumulative GPAs, and retention in undergraduate science degree programs (Haeger and Fresquez 2016; Hernandez et al. 2018). A common barrier to the institutionalization of URTPs is the fear of increased time-to-degree (Johnson and Stage 2018). However, results show that, although BUILD students attempted and completed more units

TABLE 4. Independent Sample *t*-Test Results Comparing BUILD and Non-BUILD Students across Academic Outcomes

Outcome	*n*	M (SD)	*n*	M (SD)	df	*t*	*p* value	Cohen's *d*
	BUILD		Non-BUILD					
Total units attempted	297	88.92 (33.14)	2,952	83.31 (39.02)	3,247	2.39	.017*	0.15
Total units earned	297	86.2 (32.64)	2,952	79.97 (37.78)	3,247	2.74	.006**	0.18
Completion rate (%)	297	97 (0.08)	2,952	96 (0.67)	3,247	1.42	.154	0.02
Cumulative GPA	297	3.42 (0.39)	2,952	3.17 (0.43)	3,247	9.45	<.001***	0.61
Time-to-degree (semesters)	92	7.54 (1.84)	871	7.9 (2.4)	962	-1.41	.159	0.17
	URM BUILD		URM Non-BUILD					
Total units attempted	63	106.2 (28.08)	424	103.92 (33.25)	642	0.62	.535	0.07
Total units earned	63	103.61 (27.93)	424	100.71 (32.15)	642	0.82	.415	0.09
Completion rate (%)	63	97.49 (0.05)	424	97.04 (0.05)	642	0.81	.421	0.11
Cumulative GPA	63	3.37 (0.36)	424	3.19 (0.39)	642	4.09	<.001***	0.48
Time-to-degree (semesters)	32	7.52 (1.88)	378	7.89 (2.52)	552	-1.11	.268	0.16
	Non-URM BUILD		Non-URM Non-BUILD					
Total units attempted	92	106.52 (26.25)	552	105.75 (30.07)	485	0.19	.847	0.03
Total units earned	92	104.02 (26.06)	552	102.81 (29.44)	485	0.31	.931	0.04
Completion rate (%)	92	97.57 (0.04)	552	97.29 (0.05)	485	0.42	.672	0.04
Cumulative GPA	92	3.54 (0.35)	552	3.29 (0.38)	485	4.85	<.001***	0.68
Time-to-degree (semesters)	61	7.56 (1.8)	493	7.9 (2.23)	408	-0.34	.399	0.17

Note: Underrepresented minority (URM) is defined as belonging to one of the following race/ethnicity categories: Black and African American, Hispanic or Latino, American Indian or Alaska Native, Native Hawaiian, Other Pacific Islander, Cambodian, Hmong, or Laotian (NSF 2019; Teranishi et al 2013).
*$p < .05$, **$p < .01$, ***$p < .001$.

TABLE 5. Academic and Graduate School Outcomes from Chi-Square Analyses, Overall and by Underrepresented Minority Status

Outcome	*n*	%	*n*	%	$\chi^2(1)$	*p* value	Cramer's ϕ
	BUILD		Non-BUILD				
Academic							
Graduation status	155	52.20	976	33.10	43.50	<.001***	0.12
Academic college change	61	20.50	602	20.40	00.00	.953	0.00
Graduate school							
Application	130	75.00	127	42.00	48.51	<.001***	0.32
Admission	103	79.0	77	61.00	10.59	<.001***	0.20
Matriculation	89	90.8	71	93.40	00.39	.053	-0.05
PhD programs							
Application	100	97.00	25	68.00	24.80	<.001***	0.42
Admission	85	85.00	18	72.00	02.33	.013*	0.14
Matriculation	73	89.00	17	94.00	00.48	.488	-0.07
	BUILD URM		Non-BUILD URM				
Academic							
Graduation status	178	51.70	1,731	31.90	28.30	<.001***	0.12
Academic college change	178	21.90	1,731	20.00	00.37	.543	0.01
Graduate school							
Application	83	78.90	81	46.60	28.54	<.001***	0.32
Admission	64	77.10	47	58.00	06.83	.010*	0.20
Matriculation	98	88.50	44	95.70	01.27	.188	-0.13
PhD programs							
Application	64	97.00	15	65.00	17.24	<.001***	0.02
Admission	54	84.40	12	80.00	00.17	.681	0.05
Matriculation	46	86.80	12	100.00	01.78	.183	-0.17
	BUILD Non-URM		Non-BUILD Non-URM				
Academic							
Graduation status	119	52.90	1,221	34.70	15.55	<.001***	0.11
Academic college change	119	18.50	1,221	21.00	00.41	.524	-0.17
Graduate school							
Application	47	69.10	46	35.90	19.61	.000***	0.32
Admission	39	83.00	30	65.20	03.83	.050	0.20
Matriculation	35	94.60	27	90.00	00.51	.477	0.09
PhD programs							
Application	36	97.30	10	71.40	07.69	.006	0.39
Admission	31	86.10	6	60.00	03.39	.066	0.27
Matriculation	27	93.10	5	83.30	00.60	.436	0.13

Note: Underrepresented minority (URM) is defined as belonging to one of the following race/ethnicity categories: Black and African American, Hispanic or Latino, American Indian or Alaska Native, Native Hawaiian, Other Pacific Islander, Cambodian, Hmong, or Laotian (NSF 2019; Teranishi et al 2013).
*$p < .05$, **$p < .01$, ***$p < .001$.

than their matched non-BUILD counterparts, their time-to-degree was not significantly increased, an encouraging finding for those seeking support for further institutionalization and dissemination efforts.

Additionally, BUILD students had higher cumulative GPAs and graduation rates than non-BUILD students. These results are particularly noteworthy given current mandates (Boggs 2018; Dougherty et al. 2014; LAO

2007) to increase four-year graduation rates and potential concerns that adding URTPs and coursework may delay time-to-degree (Johnson and Stage 2018). Additionally, this is one of the few studies to demonstrate that formal URTPs that serve various student divisions (first-year to fourth-year students) have a strong and positive impact on academic outcomes, including increased graduation rates. Most notably, BUILD students' GPAs remained significantly higher than non-BUILD counterparts' across URGs, suggesting that the program is helpful for many different student populations that may not typically benefit from program support.

Overall, these findings may be due to the level of social engagement within the program through yearlong faculty- and peer-led learning communities (Abeywardana et al. 2020), complemented by faculty-mentored research experiences and research-infused courses across several health-related disciplines (Urizar et al. 2017). These programmatic components engage a wider student pool interested in health research and have been shown to be high-impact practices that mainly benefit URGs in navigating social and academic cultural practices toward degree completion (Johnson and Stage 2018; Kilgo, Ezell Sheets, and Pascarella 2015).

Graduate School Outcomes

Results also showed that BUILD students, regardless of URG status, had higher application and admission rates to graduate programs, but no difference in matriculation rates compared to non-BUILD students. Additionally, more BUILD students applied to doctoral programs (97 percent) than non-BUILD students (68 percent), although their PhD admission and matriculation rates did not significantly differ. Notably, of BUILD students who applied to doctoral programs, 85 percent were admitted, and of those, 89 percent matriculated. Previous studies found students participating in URTPs to have 30 percent doctoral matriculation rates (Hall 2017; Junge et al. 2010). Together, these results demonstrate BUILD to have a strong impact on preparing students for graduate school admission and support the need for URTPs that train and prepare students for graduate school.

Limitations and Strengths

Several limitations merit mention. First, not all URTPs at CSULB were willing or able to provide participant rosters. Thus, in some instances, unless a non-BUILD student self-reported participation in research using the online survey, it is possible that URTP participation in this group was underreported. Although this may have limited the pool of possible matches, there was still a larger sample than was required. Second, PSM analyses only included students with complete data, possibly leading to nonresponse bias (Porter and Whitcomb 2005). Therefore, study findings may not be representative of all non-BUILD students at CSULB. Finally, at the time of this study, only a subset of BUILD students had graduated, limiting analysis of graduate school outcomes. Nevertheless, results for graduate school outcomes showed medium to large effect sizes (Cramer's ϕ of 0.2–0.42). Given that the BUILD Initiative is still in its pilot phase, these findings indicate that the program is effective and is contributing to students' academic and graduate school success.

Despite these limitations, this study has many strengths. First, this study serves as a model for how PSM can be used to identify a comparison group at a single institution to test URTP effectiveness across several student outcomes. The advantage of PSM is its ability to create a comparison group of students with similar demographics who are exposed to the same campus environment, activities, policies, and procedures. This comparison is key, allowing researchers to reduce the impact of confounds commonly found when comparing student outcomes across multiple campuses, institution types, and demographics. Second, multiple sources were used to obtain complete student records (i.e., institutional data, program data, faculty mentors, and student self-report), thereby increasing the level of reliability and validity of these results beyond self-report and speculation, and reducing the need to remove students from analyses due to missing data.

Future Research

The data-related challenges to testing the effectiveness of URTPs using a representative comparison group are not unique to this institution. Other researchers will likely encounter similar difficulties, limiting the strength and implications of their findings. However, results from this study support the use of PSM to control selection bias and more accurately assess program effectiveness.

Future studies should consider disaggregating health-related disciplines instead of studying BSE/BHS and STEM students in the aggregate (Haeger et al. 2020; Sax and Newhouse 2018). With larger samples, program outcomes can be evaluated by more specific participant characteristics (e.g., ethnicity, Pell grant recipients, first-generation status, gender, and intersections of race and gender). Further, there are factors beyond student preparation and training that play an essential role in degree completion and graduate school admission and matriculation. Traditionally, standardized test scores and holistic reviews of applications are part of admissions decisions, but that may be changing. The admission process relies on identifying a "fit" between the student and the program (Sowell, Allum, and Okahana 2015). Future work should focus on interventions that address nuances in admission processes and other factors that influence matriculation and degree completion.

Conclusion

This study aimed to share one method of strengthening research and evaluation practices by illustrating how to identify a comparison group using PSM for analysis of academic and graduate school outcomes. Importantly, matched comparison groups allow programs to test outcomes overall, and whether all students are being served equitably. Although significant outcomes from a one-to-one match are dependent on sample size and available data, studying the programmatic impact with a small sample was not a deterrent. Both statistical and practical significant findings can be gleaned from this work, which outweigh the time needed to collect and prepare these analyses. Overall, PSM strengthened efforts to study the impact of URTP participation on measures of academic and graduate school outcomes, allowing for more robust findings relative to the sample population.

Decades of research illustrate a need for increased funding and support for URTPs. With funding now available, and programs being piloted, effective assessment will continue to require improved efforts to compare groups of participating and nonparticipating students and program effectiveness. The use of PSM to identify statistically similar comparison groups in higher education may result in a better understanding of programmatic impact on measures of student success.

Data Availability Statement

Data from this study are not publicly available due to NIH regulations and because these data contain information that could compromise the privacy of research participants. Those interested in processes for acquiring these data can contact the authors or the BUILD program directly.

Acknowledgments

The authors acknowledge the invaluable work by Ashley Colbern in developing the BUILD SPSS database. We would like to thank the CSULB BUILD Initiative for their financial contribution, data collection efforts, and time. Additionally, we thank the California State University, Long Beach campus community, including the various departments and programs that provided data for this study. BUILD is supported by the National Institute of General Medical Sciences of the National Institutes of Health Awards UL1GM118979, TL4GM118980, and RL5GM118978. This content is solely the responsibility of the authors and is not necessarily the official views of the NIH.

Conflict of Interest

All authors were once employed by the CSULB Research Foundation and worked with the BUILD Initiative. We transparently disclose our involvement working on the BUILD evaluation and BUILD funding sources to account for any perceived conflict of interest or bias.

References

Abeywardana, Sewwandi Udeshika, Sarah Velasco, Nancy Hall, Jesse Dillon, and Chi-Ah Chun. 2020. "Near-Peer Mentoring in an Undergraduate Research Training Program at a Large Master's Comprehensive Institution: The Case of CSULB BUILD." *Understanding Interventions* 11(1) The Use and Impact of NIH-Fueled Resources for Mentoring: Reports from the Field, 12477.

Angrist, Joshua, and Jorn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Boggs, Bennett G. 2018. "A Legislator's Toolkit for the New World of Higher Education." Paper no. 4. National Conference of State Legislatures.

Brownell, Sara E., Daria S. Hekmat-Scafe, Veena Singla, Patricia Chandler Seawell, Jamie F. Conklin Imam, Sarah L. Eddy, Tim Stearns, and Martha S. Cyert. 2015. "A High-Enrollment Course-Based Undergraduate Research Experience Improves Student Conceptions of Scientific Thinking and Ability to Interpret Data." *CBE–Life Sciences Education* 14(2): ar21.

Caliendo, Marco, and Sabine Kopeinig. 2008. "Some Practical Guidance for the Implementation of Propensity Score Matching." *Journal of Economic Surveys* 22: 31–72.

Cataldi, Emily F., Christopher T. Bennett, and Xianglei Chen. 2018. "First-Generation Students: College Access, Persistence and Postbachelor's Outcomes." *Stats In Brief*. US Department of Education, NCES 2018-421. https://files.eric.ed.gov/fulltext/ED580935.pdf

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. New York: Routledge.

Cook, Thomas D. 2001. "Sciencephobia: Why Education Rejects Randomized Experiments." *Education Next* 1(3): 62–68. Accessed September 1, 2020. https://www.educationnext.org/sciencephobia

Dougherty, Kevin J., Sosanya M. Jones, Hana Lahr, Rebecca S. Natow, Lara Pheatt, and Vikash Reddy. 2014. "Performance Funding for Higher Education: Forms, Origins, Impacts, and Futures." *Annals of the American Academy of Political and Social Science* 655(1): 163–84.

Eagan, M. Kevin, Jr., Sylvia Hurtado, Mitchell J. Chang, Gina A. Garcia, Felisha A. Herrera, and Juan C. Garibay. 2013. "Making a Difference in Science Education: The Impact of Undergraduate Research Programs." *American Educational Research Journal* 50: 683–713.

Haeger, Heather, John E. Banks, Camille Smith, and Monique Armstrong-Land. 2020. "What We Know and What We Need to Know about Undergraduate Research." *Scholarship and Practice of Undergraduate Research* 3(4): 62–69.

Haeger, Heather, and Carla Fresquez. 2016. "Mentoring for Inclusion: The Impact of Mentoring on Undergraduate Researchers in the Sciences." *CBE–Life Sciences Education* 15(3): ar36.

Hall, Alison K. 2017. "Educational Outcomes from MARC Undergraduate Student Research Training." In *Diversity in the Scientific Community*, vol. 2, *Perspectives and Exemplary Programs*, 3–11. Washington, DC: American Chemical Society.

Hernandez, Paul R., Anna Woodcock, Mica Estrada, and Wesley P. Schultz. 2018. "Undergraduate Research Experiences Broaden Diversity in the Scientific Workforce." *BioScience* 68: 204–11.

Johnson, Sarah Randall, and Frances King Stage. 2018. "Academic Engagement and Student Success: Do High-Impact Practices Mean Higher Graduation Rates?" *Journal of Higher Education* 89: 753–81.

Jones, Melanie T., Amy E. L. Barlow, and Merna Villarejo. 2010. "Importance of Undergraduate Research for Minority Persistence and Achievement in Biology." *Journal of Higher Education* 81: 82–115.

Junge, Benjamin, Catherine Quiñones, Jakub Kakietek, Daniel Teodorescu, and Pat Marsteller. 2010. "Promoting Undergraduate Interest, Preparedness, and Professional Pursuit in the Sciences: An Outcomes Evaluation of the SURE Program at Emory University." *CBE–Life Sciences Education* 9: 119–32.

Kilgo, Cindy A., Jessica K. Ezell Sheets, and Ernest T. Pascarella. 2015. "The Link between High-Impact Practices and Student Learning." *Higher Education* 69: 509–25.

Kinkel, Doreen H., and Scott E. Henke. 2006. "Impact of Undergraduate Research on Academic Performance, Educational Planning, and Career Development." *Journal of Natural Resources and Life Sciences Education* 35: 194–201.

Lane, Forrest C., Yen M. To, Kyna Shelley, and Robin K. Henson. 2012. "An Illustrative Example of Propensity Score Matching with Education Research." *Career and Technical Education Research* 37: 187–212.

Legislative Analyst's Office (LAO). 2007. "Analysis of the 2007–08 Budget Bill: Education." California Legislature's Nonpartisan Fiscal and Policy Advisor. Accessed September 1, 2020. https://lao.ca.gov/Publications/Detail/1563

Linn, Marcia C., Erin Palmer, Anne Baranger, Elizabeth Gerard, and Elisa Stone. 2015. "Undergraduate Research Experiences: Impacts and Opportunities." *Science* 347: 1261757.

National Center for Education Statistics (NCES). 2001. *Education Statistics Quarterly*. Alexandria, VA: National Center for Education Statistics.

National Science Board (NSB). 2012. *Science and Engineering Indicators Digest: 2012*. Arlington VA: National Science Foundation.

National Science Foundation (NSF), National Center for Science and Engineering Statistics. 2019. *Women, Minorities, and Persons with Disabilities in Science and Engineering: 2019*. Special Report NSF 19-304. https://ncses.nsf.gov/pubs/nsf19304

Nelson, Donna J. 2008. "Nelson Diversity Surveys." Diversity in Science Association. Accessed February 20, 2022. http://drdonnajnelson.oucreate.com//diversity/top50.html

Porter, Stephen R., and Michael E. Whitcomb. 2005. "Non-Response in Student Surveys: The Role of Demographics, Engagement and Personality." *Research in Higher Education* 46: 127–52.

Rosenbaum, Paul R., and Donald B. Rubin. 1985. "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score." *American Statistician* 39: 33–38.

Rosinger, Kelly Ochs, and Karly S. Ford. 2019. "Pell Grant versus Income Data in Postsecondary Research." *Educational Researcher* 48: 309–15.

Russell, Susan H., Mary P. Hancock, and James McCullough. 2007. "The Pipeline: Benefits of Undergraduate Research Experiences." *Science* 316: 548–49.

Sax, Linda J., and Kaitlin N.S. Newhouse. 2018. "Disciplinary Field Specificity and Variation in the STEM Gender Gap." *New Directions for Institutional Research* 2018(179): 45–71.

Sowell, Robert, Jeff Allum, and Hironao Okahana. 2015. *Doctoral Initiative on Minority Attrition and Completion*. Washington, DC: Council of Graduate Schools.

Stuart, Elizabeth A., and Donald B. Rubin. 2008. "Best Practices in Quasi–Experimental Designs: Matching Methods for Causal Inference." In *Best Practices in Quantitative Methods*, edited by Jason Osborne, 155–76. Thousand Oaks, CA: SAGE.

Teranishi, Robert, Libby Lok, and Bach Mai Dolly Nguyen. 2013. *iCount: A Data Quality Movement for Asian Americans and Pacific Islanders in Higher Education*. Los Angeles: Educational Testing Service and National Commission on Asian American and Pacific Islander Research in Education. https://www.immigrationresearch.org/report/other/icount-data-quality-movement-asian-americans-and-pacific-islanders-higher-education

Urizar, Guido G., Jr., Laura Henriques, Chi-Ah Chun, Paul Buonora, Kim-Phuong L. Vu, Gino Galvez, and Laura Kingsford. 2017. "Advancing Research Opportunities and Promoting Pathways in Graduate Education: A Systemic Approach to BUILD Training at California State University, Long Beach (CSULB)." *BMC Proceedings* 11(12): 27–40.

Weston, Timothy J., and Sandra L. Laursen. 2015. "The Undergraduate Research Student Self-Assessment (URSSA): Validation for Use in Program Evaluation." *CBE–Life Sciences Education* 14(3): ar33.

Wilson, Alan E., Jenna L. Pollock, Ian Billick, Carmen Domingo, Edna G. Fernandez-Figueroa, Eric S. Nagy, Todd D. Steury, and Adam Summers. 2018. "Assessing Science Training Programs: Structured Undergraduate Research Programs Make a Difference." *BioScience* 68: 529–34.

Young, Kelly A., and Kaitlyn N. Stormes. 2020. "The BUILD Mentor Community at CSULB: A Mentor Training Program Designed to Enhance Mentoring Skills in Experienced Mentors." *Understanding Interventions* 11(1) The Use and Impact of NIH-fueled Resources for Mentoring: Reports from the Field, 12482.

Kaitlyn N. Stormes
California State University, Los Angeles, kstormes@ucla.edu

Kaitlyn N. Stormes is now at the Department of Education, University of California Los Angeles (UCLA). She is currently a doctoral student in the Higher Education and Organizational Change program at UCLA, where her research focuses on factors that facilitate or impede major persistence, retention, and graduation for students

minoritized by their gender and/or race/ethnicity in science, engineering, technology and mathematics. Stormes previously served as a senior data manager for the Building Infrastructure Leading to Diversity (BUILD) Initiative at California State University, Long Beach (CSULB).

Nicole A. Streicker earned a bachelor's degree in business and is working on her master's degree in management and leadership from Western Governors University. She currently serves as the program manager for the California State University Long Beach BUILD Initiative. Previously, she coordinated the Chancellor's Doctoral Incentive Program at CSU, Office of the Chancellor, and served on the California Forum for Diversity in Graduate Education committee. She is passionate about supporting students and improving programmatic operations.

Graham K. Bowers is now at the Department of Psychology, UCLA. Bowers earned a master's degree in psychological research from CSULB. Bowers's research examines the effects of and relationships between trauma, stress, and coping strategies within vulnerable populations. Their research goals include taking an intersectional and interdisciplinary approach to understanding how trauma and identity relate to psychopathology and treatment in underserved populations, seeking to improve individualized care for vulnerable populations.

Perla Ayala is an assistant professor and undergraduate adviser in the biomedical engineering department at CSULB. Ayala obtained her PhD in bioengineering from the University of California, Berkeley, and the University of California, San Francisco, in 2011. Ayala's research focuses on developing therapeutic systems that promote optimal healing. She serves as a training director for the BUILD Initiative at CSULB.

Guido G. Urizar Jr. is a professor of psychology at CSULB. He earned his PhD in clinical and health psychology from the University of Florida. Urizar is the director of the PRO-Health Lab (Partners in Research and Outreach for Health) in the psychology department. He is the principal investigator for phase I of the BUILD Initiative at CSULB.