Research in Higher Education (2021) 62:1248–1275
https://doi.org/10.1007/s11162-021-09632-0

SHORT PAPER/NOTE

Improving Institutional Evaluation Methods: Comparing Three Evaluations Using PSM, Exact and Coarsened Exact Matching

Bob Blankenberger[1] · **Sophia Gehlhausen Anderson**[2,3] · **Eric Lichtenberger**[2]

Received: 2 September 2019 / Accepted: 13 March 2021 / Published online: 10 April 2021

Abstract
Policymakers and institutional leaders in higher education too often make decisions based on descriptive data analyses or even anecdote when better analysis options could produce more nuanced and more valuable results. Employing the setting of higher education program evaluation at a midwestern regional public university, for this study we compared analysis approaches using basic descriptive analyses, regression, standard propensity score matching (PSM), and a mixture of PSM with continuous variables, coarsened exact matching, and exact matching on categorical variables. We used three examples of program evaluations: a freshman seminar, an upper division general education program intended to improve cultural awareness and respect for diverse groups, and multiple living learning communities. We describe how these evaluations were conducted, compare the different results for each type of method employed, and discuss the strengths and weaknesses of each in the context of program evaluation.

Keywords Propensity score matching · Exact matching · Coarsened exact matching · Program evaluation · Assessment

Higher education institutions are facing increasing pressure from accrediting bodies, and from state and federal regulatory agencies to improve accountability by creating meaningful evaluations of student success. These bodies expect institutions to provide evidence of student retention and completion (Blankenberger & Williams, 2020; Conner & Rabovsky, 2011; Heller, 2001; Ochoa, 2011; Russell, 2011; Spellings, 2006; U.S Department of Education, 2011), and of student learning (Blankenberger et al., 2017, 2020, Astin, 2012; Kuh & Ikenberry, 2009; National Governors' Association, 1986; Spellings, 2006). However, measuring student performance in meaningful ways is challenging. There are so many

✉ Bob Blankenberger
rblan2@uis.edu

[1] Public Administration, University of Illinois at Springfield, One University Plaza, PAC 421, Springfield, IL 62703, USA

[2] Illinois Board of Higher Education, Springfield, USA

[3] University of Illinois at Springfield, Springfield, USA

factors affecting student performance that isolating the impact of participation in specific institutional programs is very difficult (Perna & Thomas, 2006). Higher education administrators and policymakers too often make decisions based on descriptive data analysis or even anecdote when better analysis options could produce more nuanced, and more actionable results. With the advance of statistical matching techniques researchers have greater opportunities to do just that. In this article, we compare analysis techniques that we have employed at a midwestern regional institution to improve program evaluation processes.

Assessment of Student Outcomes as Program Evaluation

Evaluating student outcomes is not easy, but it can help to conceptualize the process as a type of program evaluation (Blankenberger & Cantrell-Bruce, 2016; Blankenberger et al., 2017; Blankenberger, 2020; Cantrell-Bruce & Blankenberger, 2015). The classic systems theory model has been in use for decades in organizational theory and provides a model to aid those engaged in evaluations. The model emphasizes the inputs, processes, and outputs of a program, within the context of its environment (Rossi et al., 2003; Sylvia & Sylvia, 2012). Generally speaking, for an educational program the core inputs include the faculty, students, and the financial, administrative, material, and technological resources necessary to deliver an educational program. As these inputs are fed into a program's educational processes, there are several expected outputs to be produced such as student learning, retention, and degree completion, as well as broader objectives such as informed citizenship, appreciation for diversity, research productivity, et al. Moreover, evaluators need to consider the many environmental and personal factors that impact student success when evaluating programs.

As Perna and Thomas (2006) note in their conceptual framework for analyzing student outcomes, evaluation of student success should be contextualized within multiple layers including internal student context, family context, school context, and the social, economic, and policy context. Accounting for the numerous factors that can impact student and program outcomes can be daunting when designing an evaluation. If evaluators were able to employ randomized experimental design this would help to account for these factors. However, as with other types of program evaluation, the ideal of experimental design can rarely be employed when evaluating educational programs.

Non-Experimental, Experimental and Quasi-Experimental Strategies

Researchers often employ a between-subjects design to analyze the treatment effects of an educational intervention by comparing the outcomes of two different groups, or a within-subjects design to examine differences between the same group pre-post treatment (Gravezetter & Forzano, 2009). Ideally, to determine whether a program/treatment produced the desired outcome, an evaluator would employ an experimental research strategy which would include an independent variable that would be manipulated by the researcher, i.e., two or more treatment conditions, a dependent variable that changes based on the treatment, and participants would be randomly assigned to experimental and control conditions. Additionally, the researcher would need to control for other covariates which could impact the relationship between the independent and the dependent variables (Gravezetter & Forzano, 2009). However, typically it is not possible to satisfy these conditions in program evaluations because

either individuals cannot be randomly assigned to treatment and control groups, because the program has already been concluded or cannot ethically be denied to individuals, or because there is no way to effectively isolate the experiment from important extraneous factors which could impact the relationship between the independent and dependent variables (Gravetter & Forzano, 2009; Rossi et al., 2003; Sylvia & Sylvia, 2012). This is typically the case in higher education program evaluations because most of the time students cannot be randomly assigned to control and treatment groups, they are not in controlled experimental settings, and they are impacted by many internal and external factors that can impact their performance (Perna & Thomas, 2006).

Recognizing these limitations, researchers evaluating education programs should seek to achieve the strongest design possible to produce results that will allow administrators and policymakers to make better, more informed decisions. Ultimately, evaluators may be constrained with the evaluation techniques they are able to employ, but they should not be satisfied with non-experimental designs that do little to minimize threats to validity. They should attempt to emulate experimental conditions as much as possible, by employing quasi-experimental designs that reduce threats to internal and external validity.

In a non-experimental educational program evaluation design, researchers may compare outcomes such as GPA or retention for two groups—either between subjects or within subjects. They then use traditional statistical techniques to compare the outcomes such as chi-square, t test, or ANOVA. If the researcher has additional data to use for control variables, they could conduct chi-square with layering, ANCOVA or regression analyses to control for the impact of some factors and determine the strength of relationships between variables. For example, in a regression equation, the researcher is able to show how much of the change in the dependent variable is accounted for by the overall model and the various factors in the model (Gravezetter & Forzano, 2009).

However, evaluation researchers can better approximate the experimental condition and improve the effectiveness of their analyses, by using quasi-experimental approaches. With propensity score matching (PSM) or exact matching the evaluator can create matched pairs to better mimic the randomization process (Austin, 2011; Thoemmes, 2012; Thoemmes & Kim, 2011). The choice of evaluation design can have important implications for the evaluation, and therefore for institutions. Although such designs cannot establish true causality, they better isolate the treatment effects associated with the intervention or program. Even the choice of whether to employ traditional PSM matching or coarsened exact matching (CEM) can have ramifications for the results of an evaluation and therefore on the choices made by those acting on those results. Researchers conducting these evaluations need to be aware of the options available to them as well as the possible impact of the design choices they make.

In this article we discuss three program evaluations we conducted at a midwestern regional public university by comparing the types of results we uncovered based on the use of traditional non-experimental techniques (t test, Chi-Square, regression), as well as quasi-experimental approaches using propensity score matching, and a mixed design combining the use of propensity score matching and exact matching. We compare the results of each type of design to show how the type and depth of the analysis can have serious implications for the evaluation and for the possible resulting actions taken by administrators and policymakers. We discuss these techniques in greater detail in the next section.

General Design and Research Questions

Like most universities, the institution in our study has attempted to increase the effectiveness of its program evaluations both to improve student and institutional success and to satisfy accreditation and government oversight expectations. The authors have conducted multiple analyses designed to measure the impact of educational and student service programs on student outcomes. However, as at most institutions, we are faced with the data access and structural limitations that accompany such evaluations. Available institutional data cannot account for many of the complex factors that impact student performance. Further, when interventions are introduced, randomly assigning students to experimental and control conditions was not an option. Therefore, to improve our evaluations, we employed propensity score matching to analyze existing datasets and imitate the randomization process in traditional experimental design as carried out by other educational researchers (e.g., An, 2013; (Blankenberger et al., 2017a; Dietrich & Lichtenberger, 2015; Gehlhausen Anderson & Blankenberger, 2020; Lichtenberger et al., 2014; Lane et al., 2012; Taylor, 2015).

Propensity Score Matching Techniques

In observational studies, propensity score matching involves applying statistical techniques to extant data to generate a propensity score which equates to the predicted probability of participation in a treatment condition to try to control for experimental bias (Austin, 2011; Thoemmes, 2012; Thoemmes & Kim, 2011). The propensity scores are utilized to produce simulated control and treatment groups that are equally likely to have participated in the treatment. Once the comparison groups are created, traditional statistical analysis techniques are employed to assess differences in outcomes between the two groups. However, since the propensity score matching process converts multiple factors, including categorical variables, into a single numeric propensity score, there is a great likelihood that the process can yield unbalanced groups on factors that may be considered important to evaluators such as race and gender (Iacus et al., 2012). Thus, some researchers have suggested that exact matching on categorical variables or coarsened exact matching, or even a mixed approach of exact matching on some variables and PSM for others may produce better results when creating matched groups (Bai & Clark, 2019; Burden et al., 2017; Imai et al., 2008; Stuart, 2010; Wells et al., 2013). Some categorical variables such as gender or the presence of a treatment may be matched exactly while others such as grade, age, or race may need to be coarsened before exact matching (e.g., coarsening several reported racial/ethnic categories to white and students of color). However, in either case, exact matching when multiple characteristics are involved can be challenging and may lead to very small matched groups which could increase the likelihood of yielding invalid findings.

These approaches provide program evaluators some valuable options, but each will have its own implications that should be considered. In this article we discuss the results of three program analyses. For each, we employed between-subjects designs, and analyzed the data with traditional analysis techniques, PSM, coarsened exact matching and exact matching. We discuss the results of each and the associated strengths and weaknesses of the techniques. Our overall research questions encompass the individual program evaluations and the overarching comparison of approaches.

(1) Controlling for academic, demographic, and non-cognitive covariates, is participation in a freshman seminar associated with greater retention and GPA?
(2) Controlling for academic, demographic, and non-cognitive covariates, is participation in engaged citizenship common experience courses associated with improved color blind racial attitude scale (racial bias measure) scores?
(3) Controlling for academic, demographic, and non-cognitive covariates, is participation in a living learning community associated with greater retention and GPA?
(4) What impacts on policy and institutional decision making can these different approaches yield?
(5) What are the comparative strengths and weaknesses of employing propensity score matching to create matched groups compared to a mix of exact matching, coarsened exact matching and PSM to create matched groups?

We will answer the first three questions in the sections dedicated to each separate study. We will consider the final two questions in the conclusion.

Methods and Results for Three Institutional Examples of Program Evaluations

Overall Approach to the Evaluations

We first conducted a simple preliminary statistical analysis, t test and chi-square, without attempting to control for covariates. Then we conducted regression analyses adding in available data on covariates to determine the extent of the relationship between the independent variable and the outcome variable while controlling for covariates. However, we were concerned that the preliminary analyses in each evaluation insufficiently controlled for possible confounders. To attempt to improve the analyses, we employed propensity score matching to create equivalent groups and control for factors that might impact the analysis of the relationship between participation in the education program and the relevant outcomes.

For all three of these evaluations, we used the propensity score matching tool in SPSS to conduct the matches. The first step was to compare the profiles of the student groups participating in the different treatment conditions (e.g., those who participated in a freshman seminar vs. those who did not) using the data available from the university's institutional research office. Running preliminary regressions enabled us to identify which factors were related to the outcome variables and which were not to determine which to include in the match. We also checked for collinearity to see which factors might be redundant.

Next, we generated the propensity scores, i.e., the predicted probabilities of student participation in a treatment group, using a logistic regression model with membership in the group as the dependent variable and the baseline attributes as the predictor variables (Austin, 2011). We included the relevant characteristics (aided by the regression and collinearity results) to generate the predicted probability, such as gender, race, GPA, et al. We then matched the group members to the nearest hundredth (i.e., a caliper of 0.01) on the key characteristics that we included in the PSM. This can be done using the PSM command in SPSS (or other statistical software program). Alternately, the propensity score can be generated by running a logistic regression with participation in the treatment/program as the dependent variable and saving the probabilities as the predictor variable for participation in

the program/treatment. The comparison groups are then matched using the predictor variable to the nearest hundredth as the cut off to create the two groups—one who participated in the treatment, one who did not. Standard t-tests or chi-squares are then run to compare the "matched groups" based on an outcome such as GPA or retention.

Although using PSM to generate matched groups improves the balance of the two groups across key characteristics, and it creates the highest number of matched pairs, this process can result in unbalanced groups. So, for each evaluation, we ran balancing diagnostics. We split the students into groups based on participation in the treatment and created output tables with the descriptive statistics. We then checked standardized differences between the two groups across each factor to use as a barometer for which factors might be unbalanced between the two groups. Although there is disagreement on the cut off score for the differences, typically, 0.2 is considered acceptable, especially with smaller groups, though 0.1 is more broadly accepted (Austin, 2011). Exceeding this indicates the groups are unbalanced on that characteristic.

However, even when matched by score this way, the groups may not match closely on key nominal variables such as race, gender or income level, particularly when there is some collinearity between two factors (e.g., race and income quartile). In each of our evaluations, we chose to add a combination of PSM on ratio factors, with exact matching and coarsened exact matching for some nominal factors. In some instances, an evaluator may consider it important to match exactly on some characteristic(s) (such as gender) to ensure the groups are matched in a way important to the evaluation, or matched by groups (coarsened exact match) in which the researcher employs grouped traits, such as for income level, region, or race. We used SPSS to transform variables into groups around race, for example, so that we could create larger numbers of coarsened matches by white/students of color as opposed to breaking out by each group separately which will reduce the number of matched pairs. Lastly, we used the sort functions in SPSS to create matched pairs based on the combination of the propensity score to the nearest 0.01 and exact or coarsened matches on nominal factors we deemed important to have exactly balanced. Again, we had to create the balancing tables to ensure the balance between the two groups was within the appropriate tolerance level.

First Year Seminar Evaluation

In 2017, one of the authors conducted an evaluation of the impact of the university's first year seminar (FYS) on student GPA and retention. This evaluation was conducted in conjunction with a few others being done at the same time as a part of an initiative by the institution to make better use of data. Freshman seminars are common, offered at almost 90% of institutions (Padgett & Keup, 2011). Typically, they incorporate orientation, study skills, and/or academic programming designed to improve student academic success. Permzadian and Credé (2016) conducted a meta-analysis of the literature findings on first year seminars finding that overall, these seminars have a small positive effect on retention (Cohen's $d = 0.11$—less than 0.2 is a small effect size) but "almost no effect" on first year GPA (Cohen's $d = 0.02$). Nonetheless, they argued that these programs are worthwhile since even small percentage gains in a campus's outcomes reflect improved success for hundreds or even thousands of students.

Data and Initial Analysis

The university had introduced the first year seminar in Spring 2011, implemented it in Fall 2012, and was ready to evaluate program impacts on student success. The institutional research office provided data on freshman from Fall 2009—Spring 2016 for the analysis, so we were able to compare students who entered prior to the introduction of FYS with those after its introduction (n = 2215). We conducted several analyses to try to determine the relationship between FYS participation and two designated outcomes: retention into fall of year two and cumulative GPA in semester three. We started with a simple between group analysis comparing students who took FYS course(s) with those who did not on these outcomes (n = 2215). We conducted the analysis for all freshman, then split the analysis groups into honors and non-honors students. The preliminary results without controlling for covariates, indicated slight gains in retention and GPA for students participating in FYS vs those who did not (see Table 1). We ran an initial logistic regression with FYS participation as the independent variable, several covariates, and retention into second semester as the dependent variable to explore the relationship among these variables. The model explained only 5.2% (Nagelkerke $R^2 = 0.052$) of the variation in the dependent variable, and only high school GPA and race were significantly related to the retention, not participation in FYS. We also ran a regression with GPA in semester three as the dependent variable (adjusted $R^2 = 0.406$). High School GPA, ACT, race, gender, and the interaction variable of FYS participation/gender/race were significant.

PSM Design

As noted in the "overall approach" section, we hoped to improve the accuracy of the evaluation by employing propensity score matching. In this evaluation, participation in FYS was the independent variable and the outcomes of retention in year two and GPA in semester three were the dependent variables. After preliminary analyses we eliminated a few factors that were either shown to not to be associated through regression, or redundant through tests for collinearity. The variables used for the initial propensity score match included the characteristics of gender, race/ethnicity, high school GPA and cumulative ACT score.

Table 1 Basic descriptive analysis: first year seminar group comparisons

	N	Retained into semester 3 (%)	Mean GPA	GPA Std. deviation
All students				
No first year seminar	1132	77.7	2.712	0.9110
First year seminar	1083	78.5	2.728	0.9468
Non-honors students				
No FYS	757	74.5	2.553	0.9464
FYS	711	75.5	2.528	0.9487
Honors students				
No FYS	375	84.0	3.031	0.7387
FYS	312	85.9	3.222[a]	0.7391

[a]Indicates statistically significant differences

Table 2 Balancing tables for freshmen seminar group participation comparisons

Variables	PSM			PSM with exact match		
	Mean or count		Difference effect size	Mean or count		Difference effect size
	Control	Treated		Control	Treated	
	N = 952	N = 952		N = 801	N = 801	
HSGPA	3.3803	3.3631	− 0.03	3.3590	3.3881	0.12
ACT	22.94	22.84	− 0.02	22.71	23.17	0.05
Male	367	378	0.02	321	321	
Race						
White	586	484		503	503	
African-American	192	270		196	196	
Hispanic	69	109		68	68	
Asian	37	28		19	19	
Multi	17	32		15	15	
Other groups	51	29		N/A	N/A	

Reviewing balancing table data, we found that balance on high school GPA and ACT were strong, but there was substantial unequal distribution on race and gender (see Table 2). We decided that given the importance of the possible differential impact on outcomes by race and gender we did not want to compare unbalanced groups. So, we changed the approach to a combination of exact matching and PSM. We matched exactly on gender and race/ethnicity combined with propensity score matching based on cumulative ACT and high school GPA. With this approach, we were able to match 1602 students, about 300 fewer than with the overall PSM match.

Results: Why the Different Approaches Mattered

If we had stopped at the basic descriptive analysis (see Table 1), we would have found that retention appeared not to be affected by first year seminar participation—a non-significant 0.8 percentage point difference, 77.7% of those who did not take FYS were retained, while 78.5% of those who participated in FYS were retained in year two (n = 2215). Similarly, GPA was not significantly different (2.90 for those who did not participate in FYS, 2.94 for those who participated in FYS). Regressions controlling for multiple covariates showed that FYS participation was not significantly associated with either outcome.

Conducting the analysis with standard PSM matching on all four characteristics (N = 1904) yielded slightly different results with nearly the same GPA for those participating in FYS (2.730) versus those not (2.734), but retention rates were slightly higher for those participating in FYS (78.5%) compared to those who did not (77.6%). Thus, we see a difference compared with the "no effect" result when not matched, though even the slight positive association in retention was considered "negligible" with a Cramer's V below 0.1 (Kotrlick et al., 2011; Rea & Parker, 1992). However, white students were heavily over-represented in the control group and African American and Hispanic students were over-represented in the treatment group. Gender representation was less uneven, but we felt the differences could still be important.

Table 3 Matched comparisons: freshman year seminar (FYS) participation group comparisons after matching

	N	Retained into semester 3 (%)	Cramer's V	Mean GPA	GPA Std. Deviation	Effect size Cohen's d
Full PSM matched students						
No FYS	952	77.6		2.734	0.9081	
FYS	952	78.5	0.010	2.730	0.9462	0.004
All exact matched students						
No FYS	801	78.2		2.732	0.8731	
FYS	801	77.9	0.013	2.769	0.9520	0.041
Female students						
No FYS	480	78.1		2.924	0.8624	
FYS	480	79.8	0.026	3.031	0.8379	0.126
Male Students						
No FYS	321	78.2		2.729	0.9313	
FYS	321	75.1	0.044	2.746	0.9536	0.018
African-Am female students						
No FYS	129	76.7		2.473	0.7001	
FYS	129	80.6	0.091	2.278[a]	0.8799	0.245
African-Am male students						
No FYS	67	86.8		2.264	0.9184	
FYS	67	79.1	0.063	2.068	0.9753	0.209
Latina students						
No FYS	46	76.1		2.534	0.8233	
FYS	46	73.9	0.013	2.411	0.9413	0.139
Latino students						
No FYS	22	95.5		2.278	0.7409	
FYS	22	72.7[a]	0.055	2.364	0.9880	0.098
White female students						
No FYS	287	79.1		2.989	0.8245	
FYS	287	79.8	0.028	3.108[a]	0.7736	0.149
White male students						
No FYS	216	74.1		2.822	0.8869	
FYS	216	75.5	0.013	2.877	0.8994	0.062

[a]Indicates statistically significant differences

Conducting the analysis with exact matching on race and gender and PSM matching on ACT and high school GPA, we were able to get exact matches on key characteristics though we were not able to create as many matched pairs (n = 1,602). Our results changed as well. Overall, 78.2% of those who did not take FYS were retained in year two, while 77.9% of those who took FYS were retained (see Table 3). This is a substantively small change (three students out of 1000), but nonetheless is noteworthy for the comparison of the three models. Instead of slight improvement for those participating in FYS as indicated in the descriptive model, we had slight decline. More notably though, there were important differences by gender and racial groups which we would have missed in both the descriptive and PSM-only models. For women, there was a slight improvement in retention for those participating in FYS (79.8%) compared to those who did not (78.1%). For men there

was a decline in retention for those participating in FYS (75.1%) compared to those not participating (78.2%). Although this is a small difference (about three out of 100 students), this result indicates that the program appeared to be associated with a negative impact for male students rather than a positive one. Furthermore, when broken out by race/ethnicity, some groups who took FYS saw increases to retention, but others did not, notably Hispanic males (no FYS 95.5% vs FYS 72.7%, about 23 students per 100) and female students (no FYS 76.1% vs FYS 73.9%, only about two per 100), as well as male African American students (no FYS 86.8% vs FYS 79.1%, i.e., almost eight students per 100). All other groups saw improved retention for FYS participants versus their matched non-FYS peers. Here again, the differences were small with a Cramer's V below 0.1 indicating a "negligible association" (Kotrlick et al., 2011; Rea & Parker, 1992) so we do not want to overly exaggerate the importance. However, this result did indicate an apparent negative differential impact for students of color which was the opposite of the purpose of the program.

The findings were also mixed for GPA. Again, the mean GPA in semester three differences were essentially negligible (effect size = 0.041) (Cohen, 1988). However, we saw some differentiation by combined race and gender groups. African American women who participated in FYS had statistically significant lower GPAs (2.473 vs. 2.278 with an effect size of 0.245 indicating a small association with FYS participation), while white female students who participated in FYS had a significantly higher GPA (2.989 vs 3.108 but with a very small effect size of 0.149). Latina women and African American male students saw non-significant lower GPAs for FYS participants though the effect size for African American males indicated a small effect size. All other groups enjoyed higher GPAs.

Impact on University Policy

The results of our analysis were fairly consistent with the findings of Permzadian and Credé (2016) showing some small but important contributions to student outcomes, but the benefits were unevenly distributed across groups. If the university had relied on the basic descriptive analysis, it might have been satisfied with the small gains in retention and GPA. After all, the literature indicated the gains might be small. Even the standard PSM approach showed the same results so may not have led to changes. However, the matched pair design enabled us to identify the more uneven results across gender and race/ethnicity groups that might not otherwise have been apparent. Discovering these differences in outcomes for underrepresented students was important. The first year seminar program was intended to improve outcomes for all students and these disproportionate impacts for students of color were concerning. Not surprisingly, the university has sought to improve the program. Using the analysis results, and feedback from stakeholders, the university launched a revision process for the First Year Seminar program in the hopes of making improvements to better serve all students. Some of what has been considered and implemented include changing the focus of the seminar to better align with general education goals, adding greater diversity awareness to the content, enhancing the orientation for FYS faculty, and more closely connecting the courses to the Living Learning Community experiences.

General Education Engaged Citizenship Common Experience Evaluation

Program Background

The second evaluation we have included is for an upper division educational program in part intended to help reduce racial bias and improve cultural understanding. Many institutions include general education outcomes related to reducing racial bias, developing citizenship, improving cultural awareness, et al. (Blankenberger et al., 2017; Cohen, 2013; Furman, 2013), but assessing these is challenging. For this program analysis, one of the authors along with several co-authors conducted an analysis of the impact of the university's upper division engaged citizenship common experience (ECCE) coursework on improving racial bias (Blankenberger et al., 2017). At the university, the general education program includes ten semester hours of upper division ECCE courses: a one semester hour speakers series course, along with three courses in two of the following three categories, global awareness (GA), U.S. communities (USC), and engagement experience (mostly internships and research projects). These courses stress acquiring greater understanding of diverse cultures in the U.S. and in the world. For this analysis, we measured racial bias as one proximate indicator for ECCE outcomes. We selected the color-blind racial attitude scale (CoBRAS) as an instrument to collect student attitudes of racial bias. The CoBRAS assessment tool is intended to "examine the degree to which individuals deny, distort, and/or minimize the existence of institutional racism" (Neville et al., 2006, p. 280). The instrument is a self-report measure consisting of 20-items rated on a 6-point Likert-type scale ranging from 1 (*strongly disagree*) to 6 (*strongly agree*). The instrument (Neville et al., 2000) scores range from 20 to 120, with higher scores indicating greater racial bias.

Data and Initial Analysis

We collected CoBRAS results through an online survey and merged the results with student demographic data through the university's institutional research office. We obtained 584 usable records out of about 3000 total undergraduates. Similar to the first year seminar study, we employed a between-subjects design (Gravezetter, & Forzano, 2009) comparing two different groups, those who took ECCE courses and those who did not. We created several comparison groups to enable us to determine whether participation in certain ECCE courses was associated with better CoBRAS racial bias scores (See Tables 4, 5, 6). In the preliminary analyses, we conducted t-tests and ANOVAs to compare the mean CoBRAS racial bias scores of different groups, such as those who had taken an ECCE U.S. communities course versus those who had not, those who had taken an ECCE global awareness course versus those who had not, and those who had completed three of their required ECCE courses versus those who had not. Several of the group differences were significant indicating an apparent benefit to taking ECCE courses (See Table 4), but with no covariates to control for group differences these are dubious findings.

Following the preliminary analyses, we performed a multivariate regression analysis to control for the impact of covariates. Several variables were significantly associated with change in CoBRAS racial bias scores, including passing a global awareness course ($\beta = -0.193$, $p = 0.007$) and passing a US communities course ($\beta = -0.154$, $p = 0.021$) as well as being a social science/humanities major ($\beta = -0.228$, $p = 0.014$), gender ($\beta = 0.147$, $p = 0.012$) and race/ethnicity ($\beta = 0.397$, $p = 0.000$) (Author, 2017).

Table 4 Group comparisons with no matching or covariates

Student course taking (comparison groups coded)	N	Mean CoBRAS score	Std. deviation	Statistical significance	Effect size Cohen's d
No ECCE	176	61.75	15.646		
Passed at least one ECCE	408	59.55	18.94	p = 0.146	
Some or no ECCE	434	61.29	17.445		
Passed GA, USC, and speakers series	150	57.11	19.337	p = 0.020*	0.227
No GA or USC course	198	62.34	15.85	No GA/USC vs	
Passed GA only	112	60.13	18.399	passed both are	
Passed USC only	98	61.57	18.772	significant at p = 0.050	
Passed both GA and USC	176	57.13	19.336	Overall ANOVA p = 0.037	
No GA course	296	62.08	16.845		
Passed GA course	288	58.30	19.002	p = 0.011*	0.211
Passed neither GA nor USC course	198	62.34	15.85		
Passed GA but no USC course	112	60.13	18.399	p = 0.289	
No USC	310	61.54	16.82		
Passed USC course	274	58.72	19.221	p = 0.061	
Passed neither GA nor USC course	198	62.34	15.85		
Passed USC not GA course	98	61.57	18.772	p = 0.713	
No speakers series	301	61.42	16.879		
Completed speakers series	283	58.93	19.117	p = 0.097	
No UIS humanities/social science course	433	61.29	18.044		
Passed a UIS hum/social science course	151	57.15	17.675	p = 0.015*	
Female	367	57.7	17.163		
Male	217	64.47	18.681	p = 0.000*	
Age 25 and under	430	59.31	17.347		
Age 26+	154	62.74	19.638	p = 0.043*	
Non-white	170	51.04	15.251		
White	378	64.03	17.955	p = 0.000*	

Lower scores indicate less racial bias

CoBRAS color blind racial attitude scale—lower scores indicate less racial bias, *ECCE* engaged citizenship common experience—an upper division general education curriculum including global awareness, U.S. communities, speaker series, among others, *GA* global awareness courses, *USC* U.S. communities courses, *SS* speaker series—a one credit course requiring attendance at a series of events/speakers

*Indicates comparison groups differences were statistically significant

Table 5 Balancing tables for ECCE group comparison examples

Variables	PSM			PSM with exact match		
	Mean or count		Difference effect size	Mean or count		Difference effect size
	Control	Treated		Control	Treated	
Not all three ECCE vs all three ECCE	N = 61	N = 61		N = 51	N = 51	
HSGPA	3.5973	3.5908	− 0.11	3.6088	3.6402	0.05
ACT	24.69	24.26	− 0.01	24.10	24.10	0
Female	46	43	0.11	39	39	
White/(vs.) non	37	39	0.07	33	33	
GA match, no USC	N = 33	N = 33		N = 27	N = 27	
HSGPA	3.7449	3.7784	0.05	3.8075	3.7508	0.02
ACT	24.97	25.18	0.06	25.41	25.48	− 0.13
Female	21	26	− 0.34	20	20	
White (vs. non)	23	20	− 0.19	20	20	
USC match, no GA	N = 25	N = 25		N = 23	N = 23	
HSGPA	3.4555	3.4138	− 0.07	3.4359	3.3704	0.10
ACT	22.32	21.96	− 0.10	21.52	21.83	− 0.12
Female	21	20	0.10	19	19	
White/non	12	10	− 0.02	15	15	

*Differences are statistically significant

PSM Design

Initially, we conducted a matching procedure using propensity score matching on the characteristics of gender, race/ethnicity, major, age, high school GPA and cumulative ACT score combined. The matching process resulted in unequal distribution of groups, particularly on race/ethnicity, age and gender (see Table 5). Since race, age and gender are significantly associated with CoBRAS scores, we changed our approach to a combination of exact and coarsened matching on these traits, along with propensity score matching. We used exact match on gender, a coarsened match on race (white/non-white), and age (25 or less/26+) combined with propensity score matching for prior academic ability incorporating high school GPA and cumulative ACT. The age cutoff separated those born before 1990 and those born after. We could have used the common 25+ cut off that is employed in federal education data reporting, but our data file had birthdates for age, so it was easy to simply split by the birthdate. Because of the limited overall cases, we already had to collapse or coarsen the racial groups into white/students of color, and we were concerned about diminishing the possible impact even further. However, we were also concerned that we might not be able to match many cases limiting the results of our analysis. Fortunately, with the full PSM process we matched 122 students, and were still able to match 102 when employing the mixed PSM and exact matching process. Differences in size for the global awareness and U.S. communities matched pairs were also minimal. Nevertheless, the small numbers limited the strength of the findings. Unfortunately, there were too few students in the over-25 sub-group to make the results very meaningful, and no relationships were

Table 6 Matched group design: all matched students 25 and under

	N	Mean CoBRAS Score	Std. deviation	Statistical significance	Effect size Cohen's d
All majors					
No GA or USC	43	61.14	16.342		
Both GA and USC	43	47.30	15.938	p = 0.000*	0.857
No GA	48	62.23	14.92		
Passed GA	48	50.65	17.634	p = 0.001*	0.709
No ECCE	27	64.41	16.851		
GA course only, no USC	27	62.07	22.087	p = 0.664	
No USC	73	57.51	16.884		
USC	73	50.47	15.718	p = 0.010*	0.432
No ECCE	23	58.96	14.044		
USC only, no GA	23	53.91	16.096	p = 0.264	
Just humanities/social science majors					
No GA or USC	19	53.47	12.366		
Both GA and USC	31	48.87	16.836	p = 0.273	
No GA	21	54.62	11.745		
Passed GA	32	49.53	17.169	p = 0.206	
No USC	30	51.60	14.724		
USC	48	47.06	14.842	p = 0.192	
Just science/business/math/comp science majors					
No GA or USC	17	66.82	18.832		
Both GA and USC	11	41.82	12.734	p = 0.000*	1.555
No GA	17	67.29	15.495		
Passed GA	15	52.07	19.274	p = 0.021*	0.877
No USC	34	62.41	18.168		
USC	23	56.48	16.099	p = 0.211	

*Indicates the comparison group differences were statistically significant

significant. In part this was because of missing data on the older students' records limiting our ability to successfully find matched pairs.

Results: Why the Different Approaches Mattered

The results for matched students 25 and under indicated a positive relationship between taking global awareness and U.S. communities courses and improved CoBRAS scores (See Table 6). This was especially true for science/business/math/computer science majors. We chose to compare this group of majors to social science/humanities majors since the latter were more likely to have already experienced repeated exposures to the type of course content that ECCE courses include. Students taking both a GA and a USC course had significantly lower (improved) racial bias scores (47.30, $p = 0.000$, Cohen's $d = 0.857$ indicating a large effect size) than their matched peers who did not participate in either (61.14). For science/business/math/computer science majors, there was an even larger difference (41.82 vs. 66.82 respectively; $p = 0.020$; Cohen's $d = 1.555$ indicating a very large effect size) compared to other majors (Blankenberger et al., 2017).

In this evaluation, if we had only conducted the basic descriptive analysis, we would have found evidence of small improvements in the CoBRAS racial bias scores, and the multivariate regression controlling for covariates would have shown that ECCE global awareness and U.S. communities course participation were significantly associated with improved CoBRAS scores. However, using the PSM matching techniques enabled us to uncover further depth. By creating these matched pairs based on gender, race, age, and major we were able to show that the effect size was very high for students taking both global awareness and U.S. communities courses, but even taking one course in global awareness yielded a large effect size, and one in U.S. communities displayed a moderate effect size. Further, the improvement was much stronger for students in science/business/math/computer science majors than for others.

Impact on University Policy

The results of the analysis and its impact on the ECCE program are still being considered by the university, but overall it was important to see that the program appeared to be positively associated with an important university learning outcome—reducing racial bias. Of course, the small number of cases, and the inherent limitations with using any such instrument to measure racial bias temper the results, and additional analyses with more students is needed, but these evaluation results provided some support for the benefits of the ECCE program. Although not surprising, it was also helpful to note the apparent greater role ECCE participation appeared to have for certain majors, and for White students, particularly male ones, though clearly more study is needed.

The university is presently revisiting the need for the ECCE in its current form. ECCE currently involves a 10-credit hour commitment for students which many departments feel is too much of a burden for students, especially since most of the university's undergrads are transfer students. Prior to this analysis little was known about student outcomes in the program. Without this analysis, it would be easier to dismiss the value of ECCE, so in that sense, any analysis showing that ECCE is associated with some positive outcome is important data. The descriptive analysis did indicate a small effect, and indicated small gender, major, and age, differences, as well as larger racial differences in CoBRAS racial bias scores. Furthermore, the regression indicated significance for these factors. However, the PSM analysis provided a richer level of analysis and stronger controls for the covariate factors. Knowing the apparent benefit of taking both a global awareness and a U.S. communities course, that the global awareness course is associated with a larger effect size than the

U.S. communities course, and the extremely large effect size for science/business/ math/ computer science majors is critical to the discussion about what the university does next. The latter is especially important given that the push for reducing ECCE hours is coming from those majors. Additionally, if the university decides to reduce the ECCE hours but not eliminate it, the effect size differences in the subgroups is valuable data. Again, we do need to keep in mind the limitations of the evaluation (especially the small "n") and that this is only one outcome, but to have this level of analysis has been considered more beneficial in the ongoing discussions than a purely descriptive non-experimental analysis.

Living Learning Community Evaluation

Program Background

The third evaluation is of the university's living-learning communities (LLCs). The university offers students the opportunity to participate in one of three LLCs: capital scholars honors program (CAP), a traditional honors program; necessary steps mentoring program (NS), created to assist first-generation college students; and students transitioning for academic retention and success (STARS) which is intended for students identified as academically at-risk (Gehlhausen Anderson, 2019; Gehlhausen Anderson & Blankenberger, 2020). Learning communities and living-learning communities are intended to improve student outcomes by building a sense of community and supporting students academically and socially (Inkelas & Weisman, 2003). They take different forms, but generally involve grouping targeted students together in classes, residence halls, and/or curricula in order to build a sense of community and give students the academic and social support they need (Inkelas et al., 2007; Zhao & Kuh, 2004). They have shown some positive impacts on students' experiences and academic outcomes (Inkelas & Weisman, 2003; Inkelas et al., 2007; Pasque & Murphy, 2005; Stassen, 2003; Zhao & Kuh, 2004) such as a sense of community, additional academic support, and the opportunity to interact with each other, staff, and faculty (Bean & Eaton, 2001), improving student levels of college engagement and stronger academic outcomes, sense of belonging, retention and graduation rates, and positive perceptions of college and residence hall environment (Cambridge-Williams et al., 2013; Inkelas & Weisman, 2003; Spanierman et al., 2013; Stassen, 2003; Zhao & Kuh, 2004). However, living-learning community participation has also shown mixed impacts on measures of academic achievement, retention and graduation, in particular across racial and ethnic groups (Cambridge-Williams et al., 2013; Noble et al., 2007; Pasque & Murphy, 2005; Purdie & Rosser, 2011).

Data and Initial Analysis

We conducted an evaluation of the relationship between participation in Living Learning Communities (LLCs—capital scholars honors, necessary steps mentoring, and students transitioning for academic retention and success) and student retention and college GPA (Gehlhausen Anderson, 2019; Gehlhausen Anderson & Blankenberger, 2020). Students must meet eligibility criteria, but are not required to take the program, so selection bias is an inherent validity concern. Hence, the need to use PSM to try to simulate equivalent groups is especially important, but this self-selection issue will always be a limitation.

Table 7 Matched group design: initial comparisons without covariates

	N	# Retained sem. 3	# Retained sem. 7	GPA semester 3	GPA semester 6
No LLC	208	167/208	126/208	2.893	2.912
Necessary steps	75	65/75	50/75	2.742	2.698
No LLC	208	167/208	126/208	2.893	2.912
STARS	71	55/71	32/71[a]	2.351[a]	2.353[a]
No LLC	208	167/208	126/208	2.893	2.912
Necessary steps and STARS	146	120/146	82/146	2.549[a]	2.528[a]
No LLC	208	167/208	126/208	2.893	2.912
Honors	223	192/223	177/223[a]	3.352	3.358[a]

LLC living learning communities—support programs that group targeted students together in classes, residence halls to build a sense of community and give students academic and social support

NS necessary steps—created to assist first-generation college students

STARS students transitioning for academic retention and success—intended for students identified as academically at-risk

CAP capital scholars honors program—a traditional honors program

[a]Indicates statistically significant differences at the 0.05 level

Additionally, the numbers are limited since historically, the university was created as an upper division institution and only admitted freshmen starting in 2001. The numbers of freshmen are still low, only about 300 of 3000 undergraduates.

We employed between-subjects design for the analysis (Gravezetter, & Forzano, 2009), comparing LLC participants with non-participants. We used three cohorts (2013–2015) for the analysis (N = 577; No LLC = 208, Honors = 223, NS = 75, and STARS = 71). As with the other evaluations, initially we conducted a basic descriptive analysis without controlling for covariates, then conducted a regression analysis to determine which factors were related to the outcomes. We found mixed results regarding participation in an LLC and improved retention and GPA. For the preliminary analysis in our matched pair design, we used no covariates and did not create matched student pairs (n = 575). Without controlling for covariates, scores were fairly consistent with a significant difference only between Honors and no LLC students in retention into semester seven (see Table 7). GPA differences were more common. However, we knew we were comparing unlike students so there was little to gain from these results. So, we ran a series of regressions that included covariates. The binary logistic regression model for third semester retention was significant overall (Nagelkerke $R^2 = 0.183$), but not significant for participation in any of the living learning communities. Similarly, the model for seventh semester retention was significant overall (Nagelkerke $R^2 = 0.216$), but in this case participation in CAP honors was significantly associated (Exp ß = 0.51), while the other two were not. The logistic regression models for third and sixth semester GPA were both significant (adj $R^2 = 0.508$ and adj $R^2 = 0.487$ respectively). Of the two models, the only LLC significantly associated with the GPA outcomes was necessary steps participation for both GPA models (Standardized ß = 0.111 and = 0.073 respectively). So, while controlling for other variables, LLC participation appeared to have little to do with improved student outcomes. However, there were several factors that were significantly associated with the outcomes providing us key characteristics for the PSM analyses.

Table 8 Balancing tables for living learning communities group comparisons

Variables	PSM			PSM with exact match		
	Mean or count		Difference effect size	Mean or count		Difference effect size
	Control	Treated		Control	Treated	
Necessary steps comparison	N = 53	N = 53		N = 29	N = 29	
HSGPA	3.2268	3.2590	0.06	3.3929	3.2752	− 0.25
ACT	21.06	20.89	− 0.07	20.83	20.83	0
Validation score	37.09	38.21	0.15	37.79	38.83	0.17
Non-white	48	42	− 0.32	22	22	0
Female	32	37	0.20	26	26	0
Pell eligible	39	41	0.09	24	24	0
Intends to complete	45	47	0.11	28	28	0
STARS comparison	N = 20	N = 20		N = 10	N = 10	
HSGPA	3.2262	3.1334	0.16	3.2675	3.3219	0.10
ACT	20.40	19.85	0.22	20.4	20.5	0.03
Validation score	37.00	38.65	0.23	40.6	36.7	− 0.64
Non-white	15	12	0.32	7	7	0
Female	13	14	0.40	8	8	0
Pell eligible	19	11	1.00	7	7	0
Intends to complete	19	17	0.30	10	10	0
NS and STARS combined comparison	N = 54	N = 54		N = 32	N = 32	
HSGPA	3.2797	3.1765	− 0.21	3.2947	3.3071	0.02
ACT	20.69	20.93	0.10	20.81	20.81	0
Validation score	38.11	36.31	− 0.26	37.84	38.25	0.06
Non-white	37	41	0.16	23	23	0
Female	38	32	− 0.23	26	26	0
Pell eligible	34	37	− 0.12	24	24	0
Intends to complete	50	47	− 0.18	31	31	0
Honors comparison	N = 77	N = 77		N = 28	N = 28	
HSGPA	3.761	3.699	− 0.08	3.8616	3.7498	− 0.22
ACT	25.5	25.4	0.12	25.18	25.64	0.16
Validation score	37.69	37.23	− 0.66	39.46	38.93	− 0.10
Non-white	22	29	− 0.21	16	16	0
Female	51	39	0.19	9	9	0
Pell eligible	26	30	− 0.11	5	5	0
Intends to complete	67	58	− 0.30	28	28	0

LLC living learning communities—support programs that group targeted students together in classes, residence halls to build a sense of community and give students academic and social support

NS necessary steps—created to assist first-generation college students

STARS students transitioning for academic retention and success—intended for students identified as academically at-risk

CAP capital scholars honors program—a traditional honors program

Table 9 Comparison of results by matching technique: retained/graduated semesters 3 and 7

Standard PSM	N	# Retained sem. 3	Effect size Cramer's V	# Retained sem. 7	Effect size Cramer's V
No LLC	53	43/53		34/53	
Necessary steps	53	45/53	0.050	35/53	0.020
No LLC	20	19/20		12/20	
STARS	20	13/20[a]	0.018	7/20	0.113
No LLC	54	46/54		35/54	
Necessary steps and STARS	54	44/54	0.050	32/54	0.057
No LLC	77	65/77		47/77	
Honors	77	53/77[a]	0.184	45/77	0.026
PSM with coarsened and exact matching					
No LLC	29	22/29		17/29	
Necessary steps	29	28/29[a]	0.300	24/29[a]	0.265
No LLC	10	10/10		6/10	
STARS	10	7/10	0.420	5/10	0.101
No LLC	32	27/32		21/32	
Necessary steps and STARS	32	30/32	0.150	24/32	0.103
No LLC	28	25/28		22/28	
Honors	28	23/28	0.102	22/28	0.000

LLC living learning communities—support programs that group targeted students together in classes, residence halls to build a sense of community and give students academic and social support, *NS* necessary steps—created to assist first-generation college students, *STARS* students transitioning for academic retention and success—intended for students identified as academically at-risk, *CAP* capital scholars honors program—a traditional honors program

[a]Indicates statistically significant differences at the 0.05 level

PSM Design

We followed up with an analysis using propensity score matching on the characteristics of gender, race/ethnicity (white vs. non-white), Pell status (ineligible vs. eligible), intent to complete (transfer vs unsure/intend to complete), validation score [using data from a Mid-Year Student Assessment™ (Ruffalo Noel Levitz, 2015) to assess students' sense of validation after they had spent two months at the institution], high school GPA, and cumulative ACT score combined (Gehlhausen Anderson, 2019; Gehlhausen Anderson & Blankenberger, 2020). Again, we discovered that the matching process yielded unequal distributions in the comparison groups, particularly on race/ethnicity, gender and Pell eligibility, but we were able to match a much larger number of students since the pairs did not have to match exactly on the four key factors (see Table 8).

Results were mixed in the full PSM analysis (See Tables 9, 10). Chi-square tests of association showed no or even negative associations between LLC participation and retention. For both STARS and Capital Scholars Honors participants, there were negative significant differences with their matched non-participant peers for third semester retention.

Table 10 Comparison of results by matching technique: cumulative GPA semesters 3 and 6

Standard PSM	N	Mean GPA Sem. 3	Std. deviation	Effect size Cohen's d	Mean GPA Sem. 6	Std. deviation	Effect size Cohen's d
No LLC	53	2.5651	0.80802		2.6351	0.81900	
Necessary steps	53	2.6642	0.76936	0.126	2.6145	0.72619	0.027
No LLC	20	2.3805	0.63761		2.3515	0.62419	
STARS	20	2.4860	0.90950	0.134	2.4190	0.91276	0.086
No LLC	54	2.6648	0.69514		2.6854	0.71699	
Necessary steps and STARS	54	2.6338	0.74884	0.043	2.6028	0.75882	0.112
No LLC	77	3.2994	0.59215		3.3029	0.58856	
Honors	77	3.0141[b]	0.61893	0.471	2.9225[b]	0.67060	0.646
PSM with coarsened and exact matching							
No LLC	29	2.4986	0.89700		2.4907	0.86692	
Necessary steps	29	2.8345	0.62827	0.434	2.7917	0.56941	0.410
No LLC	10	2.5950	0.79849		2.5570	0.83625	
STARS	10	2.4860	0.84135	0.133	2.4760	0.82452	0.098
No LLC	32	2.4316	0.85551		2.4647	0.80898	
Necessary steps and STARS	32	2.8222[a]	0.57657	0.535	2.7981	0.55851	0.480
No LLC	28	3.4164	0.52697		3.4143	0.50932	
Honors	28	3.1854	0.65808	0.387	3.2336	0.67776	0.301

LLC living learning communities—support programs that group targeted students together in classes, residence halls to build a sense of community and give students academic and social support, *NS* necessary steps—created to assist first-generation college students, *STARS* students transitioning for academic retention and success—intended for students identified as academically at-risk, *CAP* capital scholars honors program—a traditional honors program

[a]Indicates statistically significant differences at the 0.05 level

[b]Indicates statistically significant differences at the 0.01 level

Necessary Steps students showed an unforeseen result with significantly lower seventh semester retention than non-necessary steps students. For all three significant negative associations, the Cramer's V was low indicating weak to negligible strength of association (Kotrlick et al., 2011; Rea & Parker, 1992). Similarly, participation in the LLCs was not significantly positively associated with GPA at either third or sixth semester, although there was a significant negative difference for CAP honors students for both third semester GPA (M = 3.014) compared with their non-CAP matched peers (M = 3.299), and sixth semester GPA compared to non-CAP students (M = 2.923 vs M = 3.303). The effect size indicated that the former had a weak strength of association and the latter a moderate one.

Again for this analysis, we were concerned about the uneven group scores for the nominal variables after the PSM match. These factors had been significant in the regression models and could have altered the results. For example, the control group included nearly ten more students who answered "intends to complete" (rather than transfer) compared to the CAP Honors group, nearly identical to the difference in retention numbers. Furthermore, we knew that many of the honors students begin at the university but intend to transfer to sister institutions in the system after a year or two, so we needed to account for that. Similarly, the race, gender and income factors were correlated to the student outcomes and could change the results as well, particularly given the large proportion of non-white students in two of the LLCs. So, we altered our approach to match students exactly on the categories of gender, Pell status, and intent to complete, and a coarsened match on race/ethnicity (white versus students of color), combined with using propensity score matching for validation score, and prior academic ability based on cumulative ACT and high school GPA. We were very concerned about the small number of cases limiting the strength of our results with the coarsened/exact match, but we felt it would benefit the overall analysis to balance the groups and see what outcome differences might result.

Results: Why the Different Approaches Mattered

There were several differences in the revised analysis. The most notable difference was the elimination of the negative retention result we found with the CAP Honors students. As we conjectured, it appears that the difference could simply reflect the rebalance on the "intent to complete" factor. However, the challenge of exact/coarsened matching on so many factors severely curtailed the number of matches we could make. We will need to revisit the analysis in upcoming years once we have more data available. Similarly, the statistically significant difference for Honors students on GPA evaporated. Also, Necessary Steps students had significantly better retention in semester three and in semester seven compared to their matched peers with moderate associations for each (Cramer's V of 0.300 and 0.265 respectively). None of the other matched groups saw superior performance. However, the STARS students were paired with a group of students that were all retained into the third semester. This is likely an anomaly resulting from random matching and very small numbers (N = 20), something that we were concerned might happen before beginning the analysis. This indicates a potential weakness in exact match designs that could result in a relatively small number of matched pairs. To try to improve the match result and address this anomaly, we ran a propensity score match that combined Necessary steps and STARS participants. We felt justified in this merger because the NS and STARS groups have similar demographic backgrounds. This yielded 32 matched pairs (64/575 total students) and the LLC students performed better than their matched peers in both semester three and seven, though not significantly so. The revised combined group match did not change the results

for retention, but it did yield a statistically significant difference in semester three GPA at a medium effect size (no LLC 2.432 GPA, NS and STARS combined group 2.822 GPA, Cohen's $d = 0.535$).

Ultimately, if we had stopped at the basic descriptive analysis, we would have a different result for the evaluation of the program. First, the initial group comparison was not very worthwhile since the groups in the Living Learning Communities are not typical of the overall student body. Second, regression results indicated the LLCs were not associated with the outcomes, though several other factors were. Identifying these factors was useful in creating the PSM analyses. The PSM with all covariates yielded more useful results but the Honors results were stronger after the exact/coarsened match which included the match on "intent to complete", as were the results related to GPA gains for the combined Necessary Steps and STARS groups.

Impact on University Policy

We presented results of the analysis to stakeholder groups, but the results must be considered as preliminary. The results are useful to suggest some areas for further study, but due to the limited number of records available for analysis, we plan to do additional analyses before any stronger conclusions can be reached. However, there are several important themes that we noted. First, "intent to complete" appears to be an important factor for Honors students. The Honors LLC program would have preferred more positive results related to their program but recognizing the transfer patterns of their students was valuable information. The data on reduced retention from the PSM-only model would have been especially troubling, so the more neutral effect found in the PSM with exact matching model was more positive. The PSM with exact matching model findings of the positive association of Necessary Steps with retention and higher GPA that was not uncovered in the PSM-only model was helpful, though the neutral findings related to STARS was disappointing. The administration has not made any major changes to the LLC programs related to the analysis, but it has contributed to discussions about the delivery and effectiveness of the programs.

Conclusion and Discussion

We have answered the first three research questions related to the program evaluation results themselves. We have also answered the fourth question regarding the impacts on policy and institutional decision making that result from the different approaches. In all three cases the evaluation results were different based on the techniques employed. We have summarized the results in Table 11 for easier side by side comparison. The basic descriptive non-experimental designs provided some, but very limited information about the potential effectiveness of the programs. For the analysis of the first year seminar the initial descriptive analysis revealed small improvements in retention and GPA. The regression models indicated no relationship between FYS participation and student retention or GPA. The quasi-experimental propensity score matching mixed with exact matching approach proved to be much more useful. The standard PSM approach revealed about the same results, but the exact matched paired design enabled us to identify the uneven results across race and gender that would not otherwise have been apparent. Discovering these differences in student outcomes for underrepresented students was decidedly important and has

Table 11 Comparison of the approaches and results across cases

First-year seminar and retention and GPA	Results summary
1. Descriptive non-experimental	1. Indicated small improvements in retention and GPA for FYS participants
2. Regression	2.a. Results showed no relationship between FYS participation and student retention or GPA b. Several covariates were found to be significantly associated with the outcomes—important for the matching process
3. PSM	3.a. Overall, no relationships with outcomes were apparent b. Key factors were not balanced in the standard PSM
4. PSM, exact and CEM	4.a. Overall results were about the same as the standard PSM b. We were able to identify different important results across race and gender that would not otherwise have been apparent
ECCE general education and racial bias	
1. Descriptive non-experimental	1. Showed slight improvements in the racial bias scores for those participating in ECCE courses
2. Regression	2.a. Global awareness and U.S. communities course participation were significantly associated with improved racial bias scores b. Covariates race, age, gender and major were associated with the CoBRAS scores
3. PSM	3. PSM yielded unbalanced groups on key variables potentially germane to racial bias
4. PSM, exact and CEM	4.a. Yielded significant differences and large effect sizes for students taking Global Awareness and U.S. Communities courses b. Improvement in Racial Bias scores was much stronger for science/business/math/computer science majors than for others c. Matching was more of a challenge than with PSM alone and lead to small "n" matches risking the findings
Living learning communities and retention and GPA	
1. Descriptive non-experimental	1. The analysis was not useful since the groups in the LLCs could not be appropriately compared to the overall student population
2. Regression	2.a. Only the honors LLC was associated with retention and the Necessary Steps LLC with GPA b. The analysis provided useful data on other factors associated with the student outcomes
3. PSM	3. Weak negative associations were found between participation in two of the LLCs and outcomes, but several traits were not balanced
4. PSM, exact and CEM	4.a. Exact/CEM match accounted for key traits that were not balanced, e.g., student "intent to complete" b. Necessary steps students had significantly better retention in semesters three and seven and combined NS and STARS students had significantly better GPAs with a moderate effect size c. Results are only preliminary. Matching students exactly on all four important factors led to some very small "n" matched pairs

led the university to launch a substantive change to the First Year Seminar program—one which they would not have likely launched had it not been for the more detailed analysis.

For the engaged citizenship common experience evaluation, if we had only conducted the basic descriptive analysis, we would have found evidence of slight improvements in the CoBRAS racial bias scores. Similarly, the multivariate regression with covariates showed that global awareness and U.S. communities course participation were significantly associated with improved CoBRAS racial bias scores. The regression also indicated that race, age, gender and major mattered, so using the unbalanced groups that resulted from the full-PSM analysis was not satisfactory for our study. Using the PSM, CEM and exact matching techniques enabled us to add depth to the analysis and we were able to show that the effect size was large for students taking both global awareness and U.S. communities courses, that taking even one course in global awareness was associated with a medium-large effect size, and one course in U.S. communities a small effect size. Most importantly, we found that the improvement was much stronger for students in science/business/math/ computer science majors than for others. We were concerned about the small "n" sizes for the study, and that such attitudinal results have limitations. Nonetheless, these more nuanced results have proven to be very helpful, particularly since the university is currently involved in a contentious review of the ECCE program and there had been limited data regarding the relative success of the program.

In the living learning communities evaluation, the initial non-experimental descriptive analysis was not useful since the groups in the LLCs could not be appropriately compared to the overall student population. Regression models provided good data on other factors associated with the student outcomes of retention and improved GPA, but only the CAP Honors LLC was associated with retention and the necessary steps LLC with GPA. The propensity score matching with all covariates yielded more useful results. However, the Honors results were more compelling after the exact/coarsened match was conducted since this accounted for the imbalance in the student "intent to complete" as well as other key traits that were not balanced. Similarly, in the coarsened/exact match design, the necessary steps students had significantly better retention in semesters three and seven compared to their matched peers with moderate effect sizes. However, we were unable to match many students exactly since there were four important factors on which to match. This is a severe limitation and we will need to conduct more analyses as more data become available. The university is still considering what the results mean for potential changes to the living learning communities, but if the negative results for the Honors program from the first PSM had held up in the coarsened/exact PSM analysis, this may have resulted in quicker changes. Additionally, the STARS and necessary steps program have greater reason to compare what they are doing differently since one appears to have been more successful, yet their student profiles are very similar.

As for the last research question, we have noted some comparative strengths and weaknesses between non-experimental designs, regression models, standard propensity score matching, and the mix of coarsened and exact matching and PSM to create matched groups for program evaluation. First, as expected, matched pair designs provided much more valuable information than simple unmatched group comparisons. In all three instances, the unmatched group comparisons yielded results which were not supported once the matched pair designs were conducted. Although program evaluations are often hampered by limited data access, when the data are available to control for important potential confounding factors by creating matched pairs this should be done to improve the analysis results. Although evaluators are restricted by what data are available, better program evaluation includes the factors that could impact successful achievement of identified program outcomes.

Second, regression analysis alone was sufficient to show some significant relationships between the program treatments and the outcomes as well as to some covariates, but in some instances, it missed important relationships. In the ECCE general education evaluation the regression showed essentially the same association with outcomes as the matched pairs designs. However, in the freshman seminar evaluation it showed no relationship to the outcomes—unlike in the matched pair designs. Further, in the living learning community evaluation, it indicated different results related to associations than the matched designs did. Nevertheless, we recommend running preliminary regressions in each case. These are not only useful for providing analysis of associations between the treatments and outcomes, but also offers data about which covariates are related to the outcomes, and hence gives an evaluator guidance as to which factors are key for the match process.

Third, using PSM to combine multiple ratio and nominal covariates into a single predicted probability score created an improved analysis of impacts compared to basic comparisons. PSM matched pairs yielded more useful analysis results. However, checking for balance between the groups is critical, particularly on nominal variables. Often the PSM process leaves unbalanced groups and if these are on important factors that could alter the analysis this could prove problematic.

Fourth, we found that using coarsened and/or exact matching, particularly on race and gender, but also on factors like age, income, and major was a valuable addition to standard PSM. Using PSM on larger populations may make the differences in unbalanced groups negligible, but for program evaluations with more limited data sets, sometimes it is better to mix in different matching approaches on certain characteristics. For example, in our first year seminar study, we achieved a very different understanding of the outcomes associated with the program once we had created exact matches on race and gender. Similarly, without refining the matches on age and major, we may have missed critical information on the analysis of ECCE outcomes. This observation supports what some critics of PSM have suggested about the use of exact matching on categorical variables as a better option for creating matched groups (Iacus et al., 2012; Imai et al., 2008; Stuart, 2010).

However, there are limits to the usefulness of exact and/or coarsened matching when evaluating programs with relatively small numbers of individuals receiving the treatment or invention and/or relatively small numbers of potential comparison group members. For each attempt using exact matching, we lost cases, potentially diminishing the value of our results. In some instances, we only lost about one-fifth of our cases, but in others, we lost half or more. We also found that we had to collapse or coarsen what could be important data categories when the numbers became too small. This was particularly true for race where we were able to have separate comparisons for several racial groups in the First Year Seminar study in which we had 2215 cases, but we had to collapse race to white/students of color for the other two studies. This is of course an oversimplification of the dynamics within these groups and we may be missing important differences as we found with the FYS analysis. Small numbers also increase the likelihood that one could reach a spurious result, such as our Living Learning Community analysis in which we had just 20 STARS students left after the exact match—too few to make this anything but a preliminary analysis. Lastly, adding too many exact categorical matches increases the complexity of the matching process greatly. A researcher must carefully weigh the importance of the categories they are considering for exact matches compared to the time available for the analysis. Our analysis of the Living Learning Communities with four categorical exact matches was much more time-consuming than the two categorical analyses. In program evaluation, the availability of time is often just as important a limitation as other matters.

References

An, B. P. (2013). The impact of dual enrollment on college degree attainment: Do low-SES students benefit? *Educational Evaluation and Policy Analysis, 35*(1), 57–75.

Astin, A. W. (2012). *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education*. . Rowman & Littlefield Publishers.

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

Blankenberger, B. (2020). Higher education policy environment and accountability. In D. Slagle & A. Williams (Eds.), *Practicum in public administration* (pp. 11–43). Birkdale Publishers.

Blankenberger, B., & Cantrell-Bruce, T. (2016). Nonprofit education: Evaluating and assessing the skill sets that our students learn. *Journal of Nonprofit Education and Leadership, 6*(3), 243–253.

Blankenberger, B., Lichtenberger, E., & Witt, M. A. (2017a). Dual credit participation, college selectivity, and enhanced degree attainment. *Educational Researcher, 46*(5), 259–263.

Blankenberger, B., McChesney, K. Y., Schnebly, S. M., Moranski, K. R., & Dell, H. (2017b). Measuring racial bias and general education assessment. *The Journal of General Education, 66*(1), 42–59.

Blankenberger, B. & Williams, A. (2020). COVID and the impact on higher education: The essential role of integrity and accountability. *Administrative Theory & Praxis, 42*(3), 404-423. https://doi.org/10.1080/10841806.2020.1771907.

Blankenberger, B., McChesney, K. Y., Schnebly, S. M., Moranski, K. R., & Dell, H. (2017). Measuring racial bias and general education assessment. *The Journal of General Education, 66*(1), 42–59.

Bai, H., & Clark, M. H. (2019). *Propensity score methods and applications*. . Sage Publications.

Bean, J., & Eaton, S. (2001). The psychology underlying successful retention practices. *Journal of College Student Retention, 3*(1), 73–89.

Burden, A., Roche, N., Miglio, C., Hillyer, E. V., Postma, D. S., Herings, R. M., Overbeek, J. A., Khalid, J. M., van Eickels, D., & Price, D. B. (2017). An evaluation of exact matching and propensity score methods as applied in a comparative effectiveness study of inhaled corticosteroids in asthma. *Pragmatic and Observational Research, 8*, 15.

Cambridge-Williams, T., Winsler, A., Kitsantas, A., & Bernard, E. (2013). University 100 orientation courses and living-learning communities boost academic retention and graduation via enhanced self-efficacy and self-regulated learning. *Journal of College Student Retention, 15*(2), 243–268.

Cantrell-Bruce, T., & Blankenberger, B. (2015). Seeing clearly: Measuring skill sets that address the "blurred boundaries" of nonprofit education. *Journal of Public Affairs Education, 21*(3), 367–380. https://doi.org/10.1080/15236803.2015.12002204.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2nd ed.). Lawrence Erlbaum.

Cohen, J. (2013). The general education landscape. *Journal of General Education, 62*, x–xii.

Conner, T. W., & Rabovsky, T. M. (2011). Accountability, affordability, access: A review of the recent trends in higher education policy research. *Policy Studies Journal, 39*(S1), 93–112.

Dietrich, C. C., & Lichtenberger, E. J. (2015). Using propensity score matching to test the community college penalty assumption. *The Review of Higher Education, 38*(2), 193–219.

Furman, T. (2013). Assessment of general education. *Journal of General Education, 62*(2–3), 129–136.

Gehlhausen Anderson, S. (2019). Postsecondary educational attainment: The importance of living-learning communities, validation, and non-cognitive factors in affecting student outcomes (Doctoral dissertation, University of Illinois at Springfield)

Gehlhausen Anderson, S., & Blankenberger, B. (2020). Validation and living learning communities: An evaluation case study. *Journal of College Student Retention: Research, Theory & Practice*. https://doi.org/10.1177/1521025120970934.

Gravezetter, F. J., & Forzano, L. B. (2009). *Research methods for the behavioral sciences*. (3rd ed.). Wadsworth.
Heller, D. E. (2001). *The states and public higher education policy: Affordability, access, and accountability*. . Johns Hopkins University Press.
Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis, 20*(1), 1–24.
Imai, K., King, G., & Elizabeth Stuart, E. A. (2008). Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society Series A, 171*(2), 481–502.
Inkelas, K., Daver, Z., Vogt, K., & Leonard, J. (2007). Living-learning programs and first-generation college students' academic and social transition to college. *Research in Higher Education, 48*(4), 403–434.
Inkelas, K., & Weisman, J. (2003). Different by design: An examination of student outcomes among participants in three types of living-learning programs. *Journal of College Student Development, 44*(3), 335–368.
Kotrlik, J. W., Williams, H. A., & Jabor, M. K. (2011). Reporting and interpreting effect size in quantitative agricultural education research. *Journal of Agricultural Education, 52*(1), 132–142.
Kuh, G., & Ikenberry, S. (2009). *More than you think, less than we need: Learning outcomes assessment in American higher education*. University of Illinois and Indiana University.
Lane, F., To, Y., Shelley, K., & Henson, R. (2012). An illustrative example of propensity score matching with education research. *Career and Technical Education Research, 37*(3), 187–212.
Levitz, Ruffalo Noel. (2015). *College student inventory*. Retrieved from https://www.ruffalonl.com/college-student-retention/retention-management-system-plus/college-student-inventory
Lichtenberger, E., Witt, M. A., Blankenberger, B., & Franklin, D. (2014). Dual credit/dual enrollment and data driven policy implementation. *Journal of Community College Research and Practice, 38*(11), 959–979.
National Governors' Association. (1986). *Time for results*. National Governors Association.
Neville, H. A., Lilly, R. L., Duran, G., Lee, R., & Browne, L. (2000). Construction and initial validation of the color blind racial attitudes scale (COBRAS). *Journal of Counseling Psychology, 47*, 59–70.
Neville, H., Spanierman, L., & Doan, B. T. (2006). Exploring the association between color-blind racial ideology and multicultural counseling competencies. *Cultural Diversity and Ethnic Minority Psychology, 12*(2), 275–290.
Noble, K., Flynn, N., Lee, J., & Hilton, D. (2007). Predicting successful college experiences: Evidence from a first year retention program. *Journal of College Student Retention, 9*(1), 39–60.
Ochoa, E. (2011). *Renewing the American dream: The college completion agenda*. Retrieved from http://www.whitehouse.gov/blog/2011/10/05/renewing-american-dream-college-completion-agenda.
Padgett, R. D., & Keup, J. R. (2011). *2009 National survey of first-year seminars: Ongoing efforts to support students in transition* (Research Reports on College Transitions No. 2). Columbia: University of South Carolina, National Resource Center for the First Year Experience and Students in Transition.
Pasque, P., & Murphy, R. (2005). The intersections of living-learning programs and social identity as factors of academic achievement and intellectual engagement. *Journal of College Student Development, 46*(4), 429–441.
Permzadian, V., & Credé, M. (2016). Do first-year seminars improve college grades and retention? A quantitative review of their overall effectiveness and an examination of moderators of effectiveness. *Review of Educational Research, 86*(1), 277–316.
Perna, L. W., & Thomas, S. L. (2006, July). A framework for reducing the college success gap and promoting success for all. In *National Symposium on Postsecondary Student Success*.
Purdie, J., & Rosser, V. (2011). Examining the academic performance and retention of first-year students in living-learning communities and first-year experience courses. *College Student Affairs Journal, 29*(2), 95–112.
Rea, L. M., & Parker, R. A. (1992). *Designing and conducting survey research*. . Jossey-Bass.
Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2003). *Evaluation: A systematic approach*. . Sage.
Russell, A. (2011). *A guide to major U.S. college completion initiatives* (A higher education policy brief). AASC&U. Retrieved from http://www.aascu.org/uploadedFiles/AASCU/Content/Root/PolicyAndAdvocacy/PolicyPublications/Policy_Matters/College%20Completion%20October%202011.pdf
Spellings, M. (2006). *A test of leadership: Charting the future of U.S. higher education*. U.S. Department of Education.
Spanierman, L., Soble, J., Neville, H., Aber, M., Khuri, L., & De La Rosa, B. (2013). Living learning communities and students' sense of community and belonging. *Journal of Student Affairs Research and Practice, 50*(3), 308–325.
Stassen, M. (2003). Student outcomes: The impact of varying living-learning community models. *Research in Higher Education, 44*(5), 581–613.

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

Sylvia, R. D., & Sylvia, K. M. (2012). *Program planning and evaluation for the public manager*. (4th ed.). Waveland Press Inc.

Taylor, J. (2015). Accelerating pathways to college: The (in)equitable effects of community college dual credit. *Community College Review, 43*(4), 355–379.

Thoemmes, F., (2012). *Propensity score matching in SPSS*. Retrieved from http://arxiv.org/ftp/arxiv/papers/1201/1201.6385.pdf

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

U.S. Department of Education. (2011). *College completion toolkit*. Washington, DC. Retrieved from http://www.ed.gov/sites/default/files/cc-toolkit.pdf

Wells, A. R., Hamar, B., Bradley, C., Gandy, W. M., Harrison, P. L., Sidney, J. A., Coberley, C. R., Rula, E. Y., & Pope, J. E. (2013). Exploring robust methods for evaluating treatment and comparison groups in chronic care management programs. *Population Health Management, 16*(1), 35–45.

Zhao, C., & Kuh, G. D. (2004). Adding value: Learning communities and student engagement. *Research in Higher Education, 45*(2), 115–138.