TBR WORKING PAPERS

Tennessee Board of Regents
Office of Policy & Strategy

The Effect of Service Learning Participation on College Outcomes:

An Empirical Investigation

Series on Student Engagement and High Impact Practices

December 2021
Iteration 2

Table of contents

Tables

Figures

The Effect of Service Learning Participation on College Outcomes [1]

Abstract

Motivation for the study: The Tennessee Board of Regents (TBR) is implementing and expanding a suite of high impact practices (HIP) in its community colleges. Service learning, one of such HIPs, is incorporated into general education or degree programs' requirements via credit-bearing service learning components of different durations. While service learning participation and student results are examined periodically, this study contributes to the program evaluation by conducting a series of quantitative analyses that aim to assess whether a causal relationship exists between service learning experiences and key educational outcomes of first-time freshmen at community colleges.

Objectives: To investigate whether participation in serving learning HIP affects 1) the probability of earning a college credential, transferring to university, and student departure; 2) time to graduation, university transfer, and departure; and 3) academic performance. To examine if and how these effects, if any, differ by service learning duration and frequency of participation.

Methods: Different types of propensity score analysis were used, including inverse probability of treatment weighting and matching and nonparametric regression, to estimate the effect of service learning participation on outcomes of interest in general and by duration level. In the final models, propensity scores were estimated using generalized boosted modeling. Dosage analysis was employed to examine the effect of frequency of service learning participation. The following modeling techniques were used for different outcomes: logistic and OLS regression, and event history analysis (Cox regression model).

Results: Service learning participation increases the probability of graduation and university transfer, decreases the probability of student departure, expedites progression to graduation, delays progression to university transfer and departure, and is associated with a higher final GPA. The estimated results differ by duration level and frequency of service learning experience.

Conclusion: Service learning, as implemented at TBR colleges, is an efficacious HIP, which contributes to both community college freshmen student success and institutional performance.

Keywords: high impact practices, service learning, community college, program evaluation, causal inference, propensity scores, generalized boosted modeling, dosage analysis, event history analysis

[1] The primary research team included **Alexander Gorbunov**, **Amy Moreland**, **Chris Tingle**, and **Russ Deaton**. The authors appreciate insightful contribution and suggestions from Heidi Leming and Melynda Conner. The authors express their gratitude to THEC/TSAC leadership for sharing Pell and Tennessee Promise data.

Introduction

The national completion agenda and Tennessee's *Complete College* and *Drive to 55* initiatives mandate that higher education institutions become very intentional about their policies and practices that aim to help student persist to graduation and succeed both in college and in the labor market. For community colleges, this mandate may be more difficult to implement than for more selective institutions due to their commitment to the open access and transfer mission. Nonetheless, in recent years, the colleges under the governance of the Tennessee Board of Regents (TBR) have become nationally known for a host of successful reforms and large-scale initiatives that contribute to student success and institutional performance. One of such initiatives is taking high impact practices to scale with support from Lumina Foundation and the National Association of System Heads (NASH).

TBR launched its effort to promote high impact practices across its institutions in 2015. In three phases (spring 2015 through fall 2016), the number of HIPs under development and implementation had increased from three to nine. The NASH *Taking Student Success to Scale* grant offered support in 2018 and allowed TBR to expand its effort. Currently, TBR implements thirteen HIPs: Advising, Certifications, First Year Seminars/Experience, Global Cultural Awareness, Honors Education, Learning Communities, Peer Mentoring, Service Learning, Student Employment, Study Abroad, Technology Enhanced Learning, Undergraduate Research, and Work-based Learning.[2] TBR has pledged to explore the impact these HIPs have on student academic attainment, completion, and other college and labor marker outcomes.

Following up on this promise, this study continues a series of investigations into the extent and impact of TBR high income practices on various educational and, eventually, postgraduation outcomes.[3] Specifically, it employs a variety of empirical strategies to estimate the effect that service learning participation has on completion, transfer to university, academic performance, and time to graduation and university transfer. It finds that service learning exerts positive effects on key educational outcomes.

[2] TBR High Impact Practices are described at https://www.tbr.edu/student-success/tbr-high-impact-practices.

[3] The working papers and presentations on student engagement and high impact practices are available at https://www.tbr.edu/policy-strategy/presentations-and-papers

What do we know about high impact practices?

Origin and impact of high impact practices

High impact practices (HIP) are defined as a set of evidence-based teaching and learning practices that engage students in deep learning over an extended period of time, are effective for student development and engagement, and are deemed (have been tested) to have positive impact on student success in college (Kuh, 2008; Kilgo et al, 2015; Sandeen, 2012; Johnson & Stage, 2018). These practices are also referred to as engaged learning experiences, educationally purposeful activities, effective educational practices, high-impact college experiences, promising practices, good practices, or best practices—although "good practices" predate HIPs and some authors object to using the value-laden term "best practices" (Finley & McNair, 2013; Kilgo et al., 2015; Wolniak & Engberg, 2015; Hatch, 2013; Seifert et al., 2014).

The notion of HIP has been put forth and advocated by the Association of American Colleges and Universities as effective pathways to college success and career readiness (AAC&U, 2007). Originally, AAC&U identified the following ten high impact practices: first-year seminars and experiences, common intellectual experiences, learning communities, writing-intensive courses, collaborative assignments and projects, undergraduate research, diversity/global learning, service learning and community-based learning, internships, and capstone courses and projects. Later, the list of HIPs was extended to add ePortfolios as the eleventh high impact practice (Watson et al., 2016). Many of the originally identified HIPs apply more readily to four-year institutions, and most of the literature on HIP is devoted specifically to the four-year sector (Sandeen, 2012). However, there emerges a new strand of literature on HIP in community college settings (Hatch, 2013).

The AAC&U commissioned and issued a series of influential reports investigating HIPs and their impact. The first report provided a list of effective and engaging educational practices and summarized extant research on the benefits of active, engaged, and collaborative learning (AAC&U, 2007). Using data from the National Survey of Student Engagement, Kuh (2008) examined the effect of ten HIPs on student engagement and concluded that they positively affected student learning and development. He also pointed out that underserved students have more limited access to HIP than other groups of students. Swaner and Brownell (2009) offered a thorough review of the literature on HIP outcomes for underserved students. Brownell and Swaner (2010) examined research on five educational practices: first-year seminars, learning communities, service learning, undergraduate research, and capstone experiences. Kuh and O'Donnell (2013) described five cases studies on HIP implementation and focused on the HIP impact on graduation and retention rates. Overall, the AAC&U reports demonstrated that HIPs have a positive impact on key learning outcomes, such as retention, GPA, learning gains, and completion; and some of these effects differ by student demographic or academic group (Kuh, 2008; Brownell & Swaner, 2010).

Several studies investigated the impact of select high impact practices as opposed to examining a set of HIPs, and some of this research predates the AAC&U's current definition and advocacy of HIPs (e.g., Pascarella & Terenzini, 2005; Darche & Arnold, 2004; Inkelas et al., 2006; Andrade, 2007). These studies have found positive effects of participation in *specific* HIPs on such student outcomes as personal development, learning gains, student engagement, persistence, academic attainment, (shorter) time to employment, and first-year earnings—with a reservation that findings are HIP-specific (and sometimes limited to one institutions) and may not generalize well to other settings (Kilgo et al., 2013, 2015; Hu & McCormick, 2012; Largent & Horinek, 2008; Finley & McNair, 2013; Prentice & Robison, 2010; Darche & Arnold, 2004; Goff et al., 2020; Provencher & Kassel, 2017).

Finley and McNair (2013) examined the HIP effect on student engagement for underserved students, including underrepresented minority, first-generation, low-income, and transfer students.[4] Using data from 38 institutions in California, Oregon, and Wisconsin and conducting 15 focus groups, the authors applied quantitative and qualitative techniques to examine the relationship between participation in six HIPs and students' perception of their own learning.[5] They focused on *cumulative* participation in HIPs, which was operationalized as the self-reported number of distinct HIPs that students have been exposed to. The

[4] The researchers accounted for the fact that students may belong to different underserved groups.

[5] Four measures were used for self-reported changes: engagement in activities associated with deep approaches to learning, gains in practical competence, gains in general education, and gains in personal and social development.

reference group included students who did not participate in any HIP. HIPs were found to have cumulative effect: the number of HIPs that students participated in was positively related to higher levels of perceived engagement in deep learning and self-reported gains in learning, practical competence, and personal/social development. The results also confirm prior suggestions that HIP participation may be particularly beneficial for students who are underserved within higher education, providing what the authors called the "equity effects" of HIPs. Discussing their findings, Finley and McNair (2013) note, "When considering the impact of cumulative high-impact practices, it may be helpful to examine how *repeated engagement in the same type of high-impact practice* (e.g., multiple service-learning experiences) affects students' perceptions of their learning" (p. 20—emphasis added [AG]).

Kilgo et al. (2013, 2015) reported that participation in several high impact practices was significantly related to various education outcomes. Active and collaborative learning and undergraduate research were found to be most beneficial to students: they were positively related to almost all educational outcomes of interest. The other HIP with statistically significant effects were study abroad, internship, service learning, and capstone experience (the last two exerted both positive and negative effects on different outcomes; for service learning this finding was model-dependent). These experiences demonstrated lower levels of impact, which according to the authors "could suggest that some of the high-impact practices may influence student learning in a narrower way" (p. 521). Seifert et al. (2014) found that the effects of HIP on liberal arts outcomes are moderated by students' pre-college and background characteristics and thus may have various effectiveness for different groups of students.

Provencher and Kassel (2017) used propensity score matching to examine the effect of HIP participation on retention at a private four-year Catholic liberal arts college. They found that HIP participation is a statistically significant predictor of first- and second-year retention. They report that selection bias tends to increase the estimated effect of HIP participation.

Some studies did not find statistically significant effect of HIP participation, reported inconclusive results, or found counterintuitive effects for select HIPs. Johnson and Stage (2018) examined the relationship between HIPs and four- and six-year graduation rates at four-year institutions controlling for some student characteristics. They found that despite a wide-spread use of high impact practices, there is no identifiable effect on graduation rate, or the effect is negative.

Hatch (2013) examined the relationship between community college student engagement and programmatic elements of four Structured Group Learning Experiences (SGLE), which, according to the author, group distinct practices into a particular kind of HIP.[6] He found that only a few curricular elements of SGLE were positively related to engagement scores; duration and intensity (credit-bearing status) of the program did not have significant relationships with engagement; and characteristics of college personnel and students mattered most to student engagement (interestingly, these effects were negative with one exception when administrators or department heads taught or facilitated the program). Hatch concluded that there was only *limited evidence* of the relationship of program design and variation in student engagement scores. It should be noted that the researcher used self-reported data from students and administrators from the Community College Survey of Student Engagement (CCSSE) and Community College Institutional Survey (CCIS) from the selected sample of public community colleges that participated in both surveys in the same year (2012).

Regarding the postgraduation outcomes, Wolniak and Engberg (2015) examined the influence of five HIPs—internships, out-of-class research projects, study abroad, community-based projects, and capstone—on such early career outcomes as annual earnings, work environment, job satisfaction, job commitment, and continued job-related learning and challenge. They reported that the relationships between participation in specific HIPs and these outcomes were relatively small and inconsistent, with individual experiences being associated with a single outcome. The authors also conclude that "such influence is overshadowed by field of study and securing jobs closely related to majors" (p .22). They propose to exercise caution in suggesting that positive impact of HIP on student learning will lead to career gains after graduation.

[6] Hatch (2013) identifies the following four SGLE (purported HIP): learning communities, first-year seminars, orientation, and student success courses.

Service learning as a high impact practice

In service learning, students participate in a community project as part of their regular course. More specifically, Bringle, Hatcher, and McIntosh (2006) define service learning as “a course-based, credit-bearing educational experience in which students (a) participate in an organized service activity that meets identified community needs and (b) reflect on the service activity in such a way as to gain further understanding of course content, a broader appreciation of the discipline, and an enhanced sense of personal values and civic responsibility” (p. 12). Thus, the key aspects of service learning are application of what student learn in real-world settings and reflection on their service experience in classroom settings (Sandeen, 2012).

Prior findings for service learning’s impact are mostly positive but may be inconclusive in some studies. Astin et al. (2000) examined—both quantitatively and qualitatively—how service learning and community service affect students at baccalaureate-granting colleges and universities. They found significant positive effects of course-based service/community learning on most outcome measures, from academic performance to values and personal development to plans to choose a service career or participate in service after college, with benefits being strongest for the academic outcomes. Hatch (2013) found that a specific principal component of programmatic elements called *Co-curricular and Community Activities*—which included service learning alongside with campus/community service project and activities outside the classroom—consistently showed a positive significant relationship with the following engagement outcomes: Active and Collaborative Learning, Student Effort, and Academic Challenge. This principal component was also one of the least implemented components: “typical for this sample of colleges was having none or one of these activities in their SGLE” (p. 176). Finley and McNair (2013) found that service learning participants reported higher levels of deep learning engagement and perceived gains than nonparticipants; in the focus groups, students also frequently described “real-life connections” as one of four types of activities that engage them to a high degree.

Service learning and community-based learning were also found to be positively related to the following outcomes, among others: academic attainment, retention, perceived learning gains, personal and social development, critical thinking, diversity and political awareness, civic and community engagement, and global perspective-taking (Simons & Cleary, 2006; Gallini & Moely, 2003; Astin et al, 2000; Brownell and Swaner 2010; Eyler & Giles, 2001; Engberg & Fox, 2011; Valentine et al., 2021).

Wolniak and Engberg (2015) report that participation in community-based projects is positively related to sense of learning and challenge after graduation: "having more opportunities to learn new things, face new challenges, and find their work to be more useful for society" (p.16).

In contrast, Johnson and Stage (2018) did not find any effect of service learning participation on graduation rates at public institutions. In their words, "These findings do not support our original hypothesis that 1st-year seminars, writing requirements, learning communities, and service learning would be positive predictors of graduation rates" (p. 23). Kilgo et al. (2015) found that the effects of service learning participation were mixed: in the less conservative model, service learning was positively related to student learning, and the effect size was small; in the more conservative model (i.e., controlling for other HIPs), it was weakly and *negatively* related to inclination to inquire and lifelong learning. The authors posit that the latter result does not necessarily indicate that service learning participation has a negative effect on inclination to inquire and lifelong learning and suggest that future studies examine differences in HIP administration and implementation in different institutions and contexts in order to determine the true effect of service learning and other high impact practices.

High impact practices at TBR

In the past several years, TBR has developed detailed taxonomies of its thirteen high impact practices and introduced a system goal of having all students experience two HIPs before they complete a degree at a community college. As part of the HIP quality assurance, TBR has developed a "minimum definition of practice" for each HIP that it implements. For example, for service learning—which is the focus of this study—TBR has the following minimum definition of practice: "Service-learning is a teaching and learning strategy that integrates meaningful community service with instruction and reflection to enrich the learning experience, teach civic responsibility, and strengthen communities. Curriculum includes structured field-based “experiential learning” alongside community partners, which reinforces course learning outcomes. Within the TBR System, credit-bearing service-learning designated courses are incorporated into general education or college core requirements for a degree program” (TBR, n.d.).

As part of its ongoing investigation of high impact practices, TBR has examined student participation in first-year experience and its effect on several short-term outcomes. The analysis shows that both participation in first-year seminars and the way this HIP is implemented differ by student group and college. First-year experience was not found to affect fall-to-spring or fall-to-fall retention or gateway course completion in English or mathematics. However, these results also differ based on how colleges implement first-year seminars. Future analyses will include longer-term outcomes (TBR, 2020).

Recently, Lumina Foundation commissioned a study to examine the effects of high impact practices at TBR community colleges participating in the Lumina-NASH HIP initiative (Valentine & Price, 2021). The analysis was conducted at five TBR community colleges and examined the impact of HIP participation in the first term on a set of short-term educational outcomes for three cohorts of students, fall 2017 through fall 2019. The study found large positive effects of HIP participation—including service learning—on fall-to-spring and fall-to-fall retention, credit accumulation, and completion of gatekeeper courses in English and mathematics within the first year for all students, but also for underrepresented groups. These results were especially notable for Black and adult students. Based on prior research linking short-term academic outcomes and degree completion, the analysts suppose that these positive impacts on short-term outcomes will eventually lead to higher credential completion for HIP participants. Service learning was found to have positive effect (small as compared to the impact of first-year experience and undergraduate research) for most outcomes for all students; and service learning effects were described as “large and consistent benefits across outcomes” (p. 8) for Hispanic students.

This study contributes to this body of research by addressing the selection bias that is steeped in factors driving selection into high impact practices and focusing on service learning as the primary HIP of interest. Future investigations will examine the effect of participation in other TBR high impact practices on different college and labor market outcomes as well as conditional effects for specific student groups.

Methodology

Key terms

In this section, the following terms and abbreviations are used in the context of the current investigation (Rubin, 1974; Rosenbaum & Rubin, 1983; Heckman & Robb, 1985; Austin, 2011a; Guo & Fraser, 2015; Murnane & Willet, 2011; Smith & Todd, 2001; Vogt, 2005).

Treatment (exposure) – participation in service learning experience or its specific duration level. Depending on the model, treatment can be binary (participated or did not participate in HIP, coded to "0" and "1"), or multiple/nonbinary (i.e., how many times participated in this HIP or its specific level, coded to "One time", "Two times", and "Three or more times").

Confounding variable – a variable in the model that obscures (confounds) the effect of another variable. In this study, pretreatment variables that affect both treatment assignment and the outcomes of focus are confounders: they make it harder to separate their effects from the treatment effect.

Selection bias – A problem that arises in comparison of groups when groups are formed by subjects that choose to join them instead of being randomly assigned to these groups by a researcher. Student characteristics influence selection into treatment; as a result, in observational (non-experimental) studies, the estimated treatment effect is confounded by student characteristics. Thus, service learning participants and nonparticipants are expected to be different in important ways, including their goals and motivation.

Counterfactual – A fundamentally unknowable concept of what would have happened had something occurred that did not actually occur. In this study, the counterfactual represents the outcome participants would have experienced, on average, had they not participated in service learning. It is a hypothetical (non-existent) group, and our methodology aims to create a comparison (untreated) group that looks sufficiently like the treated group as a way of approximating counterfactual conditions. All methods described in this section seek to estimate the counterfactual for service learning participants.

Propensity Score (PS) – the probability of treatment assignment (or treatment selection) conditional on baselines characteristics (a set of observed covariates that precede the treatment). The true PS is unknown; and to remove bias completely, it should contain all confounding factors. In observational studies, PS is estimated based on covariates that are deemed to affect selection into treatment. Estimated PS is a scalar index (single score) that summarizes the information that covariates used to predict treatment contain and that explains the systematic nature of selection. Propensity scores are estimated with the goal of balancing covariates between treated and untreated subjects in order to isolate the treatment effect. In this sense, PS is a balancing score, which represents a vector of covariates, and it is used to create balanced datasets. Subjects from treated and untreated groups with similar PS are analytically comparable, although they may differ on values of individual covariates.

Generalized Propensity Score (GPS) – the probability of receiving a particular level of treatment (participating in HIP a given number of times), conditional on pre-treatment covariates. GPS has a balancing property, which is similar to propensity score for binary treatments and can be estimated for multiple or continuous treatments. In this study, GPS are estimated for models that examine the effect of frequency of service learning participations.

Propensity Score Analysis (PSA) – a family of statistical techniques that are used in causal modeling when it is not possible to conduct a randomized experiment. The most common methods in PSA include propensity score matching, propensity score weighting, stratification on the propensity score, and covariate adjustment using propensity score.

Average Treatment Effect (ATE) – the effect attributable to treatment among the population. In other words, it is the effect of treatment across the entire population of treated and untreated subjects (cases). It is the average effect of moving the whole population (in this study, all freshman students) from untreated to treated. ATE is the main estimand (treatment effect of interest) in this analysis.

Average Treatment Effect for the Treated (ATT) – the effect attributable to treatment among those who underwent treatment. It is the average effect of treatment on subjects who actually received the treatment. ATT is the estimate of interest when analysts want to know if, on average, the treatment was beneficial for those who select (are assigned to) treatment, but not for the whole population. ATT is the estimand in several secondary models in the analysis (propensity score matching and kernel matching).

Research questions and identification problems

This study investigates two main research questions: First, does service learning participation affect the following outcomes: 1) the probability of earning a college credential, the probability of transferring to a four-year college or university, and the probability of student departure; 2) time to graduation, transfer, and departure; and 3) academic performance? Second, do these effects differ by service learning duration (number of hours in the component) and frequency of service learning participation?

Thus, this investigation aims to estimate the effect of service learning participation on the outcomes of interest in general, by duration level, and by frequency of participation in service learning and its individual duration levels. These research questions and comprehensiveness of the analysis presents several estimation problems that need to be addressed.

First and foremost, we expect that there is a selection bias due to students selectively choosing the treatment (service learning participation), and participants and nonparticipants being systematically different on a host of characteristics. Some of these characteristics may affect just treatment assignment, while others may affect both selection and the outcomes of interest. While some of these factors may be available for inclusion in regression adjustment models as control variables, it is also quite possible that HIP participants differ from nonparticipants on things that cannot be obtained or measured directly and reliably: motivation, aptitude, approaches to college planning and ability to follow through with it, family circumstances, peer and significant others' influence, prior education history and college intent, and so on. These observable and unobservable characteristics will also affect the outcomes of interest, making it more difficult to estimate the effects of service learning participation that may have causal interpretation. Under these conditions, regular regression methods cannot recover causal evidence of the effect, and a direct comparison of groups of subjects will not overcome the problem of identification.

We address the selectivity bias with several variations of propensity score analysis, which are described below. As a robustness check, we also employ a doubly robust estimator, which combines fitting models with inverse probability of treatment weights with the inclusion of additional pretreatment control covariates (i.e., critical demographic and academic variables) in order to minimize any remaining bias. The doubly robust estimation provides another chance to correctly specify the model and—if either the propensity score model or the multivariate outcome model is specified correctly—to obtain more consistent and unbiased estimates of the treatment effect. So, it provides additional protection against misspecification of any model and gives two chances for a valid inference (Bang & Robins, 2005).

The second key estimation issue is that data under analysis is characterized by time dependence and censoring, which must be accounted for. Students participate in service learning in different semesters, and thus participants may have varying amounts of time before they graduate, transfer, depart from higher education, or exit the study. The probability of attaining the outcomes—graduation, transfer, or departure—changes over time, students remain in the study for different lengths of time and thus provide data for different durations, some students may experience the outcome after the end of the observation. The effect of service learning participation on the outcome may change over time, and there are time-varying confounding factors (such as age or institution of enrollment). In addition, there is a problem of the so-called *tied events*: because time in the study is measured in calendar terms and many students graduate, transfer to university, or depart during the same semester, it is impossible to estimate individual durations of time and determine the order of event occurrence (outcome attainment). This issue hinders the precise estimation of time-to-outcome effects.

These time-related issues of estimation are addressed by employing the appropriate methods of Event History Analysis (EHA), which aim to estimate the effect of treatment on time to the respective outcome. We discuss the specific EHA model that was used in the study below.

Finally, there are some data issues that need to be addressed for purposes of quantitative analysis. Because the data for the study comes from different sources, some data elements may be missing for semesters in non-TBR institutions. In such cases and when the values could change over time (e.g., for college GPA), we used the last-observation-carried-forward method to impute missing data for all post-TBR terms. Also, 14,6 percent of students in the cohort do not have an ACT score reported for them. After making sure that there were no identifiable patterns in data missingness, we used multiple imputation with 15 variables and 50 imputations to address this issue.

The multitude of the estimation problems required the use of several estimation techniques and their combination in order to estimate the treatment effect. We address the selection bias by using variations of propensity score analysis (inverse probability of treatment weighting in the main models but also kernel weighting and matching in the models used for sensitivity analysis). We use the following techniques to account for different outcomes and address pending estimation issues: logistic and Ordinary Least Squares (OLS) regression and Event History Analysis (Cox regression model). For comparison, we apply these techniques both to the original (unweighted/unmatched) samples and to the weighted or matched samples used in the final models or models in the sensitivity analysis.

Approaches to estimating propensity scores

The most common way of calculating propensity scores is via a logistic or probit regression. In this approach, the treatment (exposure) variable is a binary outcome with multiple covariates that are deemed to affect the treatment selection (exposure causes) being used as predictor variables. The estimated propensity score signifies the probability of getting the treatment; in the context of this study, it is the probability of participating in service learning (or its individual duration levels). We employ this traditional approach to estimating propensity scores in some models that are used for sensitivity analyses.

There are certain disadvantages to using the above method of propensity score estimation. One cannot be certain that the functional form of the model is specified correctly. Logistic regression specifies linear relationships between predictor variables and the logit of the propensity score, and this assumption may not be met. It is highly recommended that covariates be included in the correct functional form. However, the propensity score model using logistic regression may be misspecified due to omission of non-linear or interactive terms. Misspecification may lead to not removing all confounding bias or, in the worst scenario, to increasing it. In order to address the issue of potential misspecification, alleviate bias due to omitted nonlinearities, and achieve the optimal balance, analysts are encouraged to continue respecifying the model by including various interaction terms and/or high-order polynomial terms. Also, logistic regression can produce unstable propensity scores and generate extremely large weights (Wang et al., 2019; Austin, 2011a; Thoemmes & Ong, 2015; Morgan & Todd, 2008).

Because of the above issues, for our investigation, we employ a different approach to estimating propensity scores, which is deemed superior to alternative methods of estimating sample propensity scores and predicting treatment assignment. Known as Generalized Boosted Modelling (GBM), this technique relies on machine learning algorithm (data-mining technique) in propensity score estimation. The algorithm fits multiple models, using a modern boosted regression trees approach, and then merges their predictions. According to Schonlau (2005), "There is a mounting empirical evidence that boosting is one of the best modeling approaches ever developed" (p. 331).

GMB offers numerous advantages that made it an ideal choice for our investigation. First, in contrast to logistic regression, GBM takes care of the functional form of the model and can handle all kinds of conditioning variables, their transformations, and nonlinear and interaction effects. In fact, according to McCaffrey and co-authors (2004), the adjustment of the propensity score is the same no matter what form of the predictor variable—linear, squared, logarithmically transformed, etc.—has been used. Thus, analysts do not have to specify the functional form of the covariates in the model. Second, it allows for using a large number of covariates in the propensity score model. Boosting will work even when there are more variables than observations. Third, by using random subsamples (training data), it further reduces prediction error in the estimation. Finally, it has been shown that GBM performs better than logistic regression and alternative machine learning methods in estimating propensity scores, especially with the IPTW approach (McCaffrey et al., 2004; Ridgeway, 2007; Ridgeway et al., 2020; Guo & Fraser, 2015; Lee at al., 2010; Schonlau, 2005; Harder et al., 2010; Wang et al., 2019).[7]

For the purposes of developing propensity scores, the optimal number of trees is determined by minimizing the average standardized absolute mean difference (ASAM) between the treatment and control groups (McCaffrey et al., 2004). For each propensity score model in our analysis, the algorithm was

[7] We used the *Toolkit for Weighting and Analysis of Non-equivalent Groups* developed by the RAND corporation (Cefalu et al., 2015; RAND, n.d.; McCaffrey et al., 2004).

stopped when ASAM in the covariates was minimized. In other words, for each treatment variable as the outcome of the propensity score model, we used the number of iterations that minimized the absolute standardized mean difference (effect size).[8] We also allowed for four-way interactions (four splits for each simple tree) and used a shrinkage coefficient of 0.005 for a smooth fit.

Covariates for propensity score models

Selection of variables for our propensity score model was driven by the following considerations: recommendations in extant literature on which variables to use, prior research on predictors of student outcomes and HIP participation, knowledge of service learning implementation in TBR colleges, and data availability and quality. For the first consideration, views range from including all observed baseline characteristics to using only those covariates that are deemed to affect both treatment assignment and outcomes. Such decisions have consequences for bias reduction, on the one hand, and efficiency of estimation, on the other hand. It may be a better choice, however, to include a variable that is related to treatment assignment but not the outcome (losing some efficiency) than to omit a variable that is related to both and bias the outcome (Rubin & Thomas, 1996). Several data-driven approaches have been suggested for variable selection for the propensity score model using logistic regression, from stepwise logistic regression to different statistical criteria (Rosenbaum & Rubin, 1984; Hirano & Imbens, 2001; Dehejia & Wahba, 1999).

Despite divergent views on which covariates to use, we heed the advice that it may be beneficial to include all measured covariates because, in practice, most subject-level factors likely affect both treatment assignment and the outcome (and thus reduce bias without compromising efficiency), provided that the dataset is sufficiently large (Austin, 2011a; Garrido et al., 2014; Feng et al., 2012). Moreover, for GBM it has been suggested to use all available covariates as the algorithm will adaptively choose the ones to include (McCaffrey et al., 2004). However, although we employ GMB on large longitudinal datasets, we

[8] This approach is different from the standard GBM and an alternative boosted regression model in the sensitivity analysis using Stata's *boost* program, both of which minimize prediction error and not ASAM.

still include covariates as the treatment predictors that are motivated by theory, prior research, and our knowledge of service learning participation.

In our final model, we use the total of 33 baseline covariates for propensity score estimation that fall into the following broad categories: demographic factors (age, gender, race/ethnicity groups, resident status, and being Pell-eligible at any time in educational history as a proxy for socioeconomic status), academic variables (high school GPA, ACT composite score, high school diploma type, enrollment delay, being a learning support student, receiving assistance via the Tennessee Promise program, enrollment status and attempted credit in the first term, degree intent, and enrolling in Tennessee Transfer Pathway), financial aid variables (amounts of the following grants: Pell, Tennessee Promise, Tennessee Lottery, and Tennessee Student Assistance Award), and major field categories (majors in the first term of enrollment). Based on CIP, the following eight major field groups were identified and used in the propensity score model: Applied Technology, Arts, Business, Education, Health, Humanities, Social Sciences, and STEM.

In some sensitivity analysis models, it was possible to include other, more finely grained, covariates or, on the contrary, impossible to include the same number of covariates due to estimation problems, and we had to use fewer baseline characteristics. For example, the Stata package for boosted regression allowed including 51 CIP codes and 13 institutional flags. In contrast, the models using logistic regression to predict propensity score could not handle continuous variables for financial aid and we had to represent Pell amount by five distinct categories (no other financial aid variables were used). The logistic models also used two variables for the degree intent (“terminal” and “transfer”, with “undeclared” as the reference group), flag for enrollment status, and credits earned (also one major group was used as a reference category). The total number of covariates in the logistic models was 26.

Appendix 1 and **Appendix 2** present the list of all covariates that were used in the final (GBM-based model) while also assessing the achieved balance on them after applying the inverse probability of treatment weights. **Appendix 3** graphically shows the reduction in standardized differences between the groups of service learning participants and nonparticipants due to applied ATE weights.

In our investigation of service learning impact, we follow a common practice in propensity score analysis: we employ different methods and use comparisons across models in order to draw conclusions and decide on the main model. The sections below discuss both the main (final) models and the comparison models used for sensitivity analysis.

Main model: Estimation of the treatment effect by weighting by the inverse of the propensity scores

Imbens and Wooldridge (2009) showed that if the propensity scores are estimated correctly, using the inverse probability of treatment weighting (IPTW) is an effective method to neutralize the effect of selection on average treatment effect estimate. The main idea behind this method with propensity scores is to use probability weights to control for confounding. Using weights ensures that the distribution of confounders is the same for treated and untreated groups, which effectively results in removing these confounders. This approach weights more heavily untreated subjects who are more similar to treated cases and reduces the weights of untreated units who are dissimilar. The weight is estimated as the inverse of the propensity score (the probability of receiving the treatment). IPTW creates a pseudo-population that is representative of the treated group in terms of distribution of confounding factors. Stated differently, it constructs a comparison group of untreated subjects who are observationally similar to treated cases, and in the synthetic sample, the distribution of baseline covariates is independent of treatment assignment. The IPTW is used as the main estimator in this study due to its demonstrated ability to minimize bias, achieve balance on covariates, and yield statistically efficient estimates in comparison to other propensity score methods. In addition, IPTW is recommended for longitudinal data with time-varying confounders, which are present in this investigation (McCaffrey et al., 2004; Imbens, 2004; Austin, 2011a; Austin & Stuart, 2015; Hirano et al, 2003; Thoemmes & Ong, 2015; Joffe et al., 2004; Imbens & Wooldridge, 2009; Sato, & Matsuyama, 2003; Guo & Fraser, 2015).

As McCaffrey et al. (2004) state, “For a large sample size, the weighted treatment effect estimate will be nearly unbiased provided that several assumptions hold” (p. 405). To address the selection bias with IPTW, the following assumptions must be met: there are no unobserved confounding factors (observed covariates explain all pre-existing differences affecting outcomes); each subject has a nonzero probability (but not a probability of 1) to receive treatment; and the IPTW model is specified correctly (Thoemmes & Ong, 2015; McCaffrey et al., 2004). Although the first assumption is untestable, we select covariates based on prior literature, past research at TBR, and knowledge of factors that influence participation in service learning opportunities. Thus, we argue that by modeling the relationship between observed covariates and treatment selection, we account for all key confounding factors. To make sure that IPTW helped correct for selection, we check balance on observed covariates using weighted standardized difference. Tabular and graphical assessment of balance in **Appendices 1**, **2**, and **3** demonstrate the effect of using IPTW on balance of observed covariates. The plausibility of the second assumption is assessed by examining if the empirical propensity score distributions overlap. We use boxplots of propensity scores to check for such overlap between treated and untreated subjects in the propensity score space (**Appendix 4**). Finally, by using generalized boosted regression to estimate sample propensity scores, we address the issue of potential misspecification of the IPTW model.

To summarize, we use the IPTW in our main model, with Generalized Boosted Regression modeling as the key method for estimating propensity scores. We also take two additional steps when we apply weighting to our models to address the following issues. First, any sample with weights may include atypical cases. Such extreme cases may include treated subjects with very low estimated propensity score and untreated subjects with very high estimated propensity scores. The weights for such atypical cases may be unstable and/or inaccurate (too large or two low). To avoid assigning too much weight to extreme cases, we truncate the weights at 1% and 99% of the weight distribution (Cole & Hernan, 2008). Following the advice by Imbens (2004), we also normalize the weights to one. Second, one must account for variability in the original propensity score model and the fact that weights were estimated, which could lead to an increase in variance. To address this issue, we use robust ("sandwich") standard errors to calculate the adjusted standard errors and confidence intervals for our estimates (Austin, 2011a, 2016; Thoemmes & Ong, 2015).

Secondary models in sensitivity analysis

The study underwent two iterations. In Iteration 1, the observation period was limited to nine semesters (through summer 2020). In 2021, the study was conducted again to include twelve terms of tracking data through summer 2021 (Iteration 2).

In Iteration 1, we explored the sensitivity of our estimates to different model specifications. We refer to this approach as sensitivity analysis (in a broad sense, as explained below) or comparison models. The main models included in sensitivity analysis are described below but, overall, include the following: alternative ways of estimating and applying IPTW (using a Stata package for boosted regression and traditional logistics regression), Kernel matching/weighting (with universal and treatment-specific bandwidth), propensity score matching for nearest neighbor with caliper (2, 4, and 6 neighbors), and one-to-one propensity score matching. It is critical that the IPTW models estimate the Average Treatment Effect, while kernel and propensity score *matching* models estimate the Average Treatment Effect for the Treated. Thus, it is not quite correct to refer to the latter group of models as “sensitivity analysis”; however, in this report we use this term in a loose sense to indicate specifications that were run to compare the results (statistical significance, direction of the effect, effect size) of the main models. For each sensitivity analysis model, we used the same analytic tools (logistics and OLS regressions and EHA) but relied on other approaches to calculating and using propensity scores.

This section briefly describes the comparison models in the order in which their results are presented in **Appendices 5-9**. GBR Modeling in these tables represents the main model of the study as of Iteration 1 of the investigation. The first set of models estimate the Average Treatment Effect (ATE). We checked the achieved balance on covariates both with formal tests and examination of balance graphs.

IPTW using a Stata* boost *package

As an alternative data adaptive algorithm, we ran boosted regression for propensity score estimation using a Stata *boost* package (Schonlau, 2005). The difference with the main model is that this boosting algorithm minimizes the prediction error and not the average standardized absolute mean difference. The advantage of this package is that we could include more predictor variables in propensity score estimation than in any other model that we tested. For example, the *boost* command could handle inclusion of 51 CIP codes in the model. However, we decided against using it as the main model in our analysis for two reasons. First, this package does not allow stopping the algorithm when ASAM is minimized. Second, we discovered that predicted probability of treatment (used to create inverse probability of treatment weights) was both sensitive to the order of covariates and varied with each run, which affected the replicability of results. The solution was to "freeze" the model once we saw that the balance on covariates was achieved. The produced weights were normalized and truncated.

IPTW with logistics regression

As part of the sensitivity analysis, we also created inverse probability weights using a more traditional approach—logistic regression. As explained below, the conditioning variables used in propensity score estimation were slightly different from the main model: the convergence was never achieved when we used exactly the same set of covariates. The developed weights were normalized (or standardized in an alternative specification) and truncated.

The second set of models in the sensitivity analysis group estimates the average treatment effect on the treated (Heckman & Robb, 1985; Smith & Todd, 2001). In other words, in these models we estimate the impact of service learning on outcomes for students who actually participated in this high impact practice. Namely, we use various matching techniques to estimate the effect of interest (Morgan & Harding, 2006). We start with the non-parametric kernel matching and proceed with two variations of propensity matching: nearest neighbor matching with caliper and one-to-one matching.

Kernel matching

To estimate the counterfactual, kernel matching (kernel weighting) weights each subject in the untreated group based on the distance from the given treated subject. Each treated subject is assigned a weight of one. Untreated subjects are assigned different weights; the weight is the highest for the control units that a closest to a given propensity score, but the weight rapidly approaches zero, the father away control units are from this propensity score. Better matches get larger weights, and each match for the treated subject is a weighted composite of untreated units (within a range/bandwidth). In other words, this method uses propensity scores derived from multiple matches to calculate a weighted mean that is used as a counterfactual. The distance between the target treated subject and all untreated subjects is transformed with an aid of the kernel function. Kernel weighting uses nonparametric regression, and bootstrapping must be used with to draw statistical inferences (Heckman et al., 1997, 1998; Guo & Fraser, 2015; Garrido et al., 2014; Caliendo & Kopeinig, 2008).

Propensity score matching has been a popular method in causal inference studies and has a rich literature describing its application. However, it has been argued based on simulations that propensity score matching may increase imbalance relative to the original data, lead to inefficiency and bias, and increase model dependency. All these outcomes are the opposite of the intended effect. Also, because propensity score matching approximates complete randomization, it is inferior to alternative matching methods, which approximate a more efficient fully blocked randomization and thus can achieve lower levels of imbalance and bias (King & Nielsen, 2019; King et al, 2011; King, 2015). At the same time, it has been also shown that these problems apply mostly to a specific type of matching (one-to-one matching without replacement) and manifest themselves only under certain conditions (Jann, 2017).

Nearest neighbor matching with caliper

This method of matching constructs the counterfactual for each subject in the treatment group by using the untreated subjects that are closest to the treated one on the estimated propensity score. To avoid having poor matches, we used a prespecified caliper to restrict matches to a maximum distance. We set the caliper to 0.2 of the standard deviation of the logit of propensity score based on findings from simulation studies that showed that caliper of this width (or one close to it) is optimal (Austin, 2011a; Garrido et al., 2014). We used three variants of this model: with two, four, and six nearest neighbors, while keeping ties (other untreated cases with identical propensity scores, if any) and with common support only. Imposing a common support means omitting treatment observations whose propensity score is higher than the maximum or less than the minimum propensity score of the control cases. We used Abadie and Imbens standard errors for these models.

One-to-one matching on common support

The last method of estimating the effect of service learning (as a binary treatment) on college outcomes consisted in using one-to-one match on estimated propensity score. As the name suggests, in this approach, each subject who received treatment is matched to one untreated individual, and everybody is assigned a weight of one. The resultant sample size is smaller than in the previously described methods due to omitting unmatched cases. Smaller sample size may affect statistical significance of the results as compared to alternative models. As in the previous approach with nearest neighbor matching, the caliper is set to 0.2 of the standard deviation of the linear propensity score. Matching was done with common support and without replacement. Abadie and Imbens standard errors were used. This approach also estimates the Average Treatment Effect on the Treated.

Dosage analysis for non-binary treatments

For non-binary treatments (frequency or dosage analysis), we also implemented propensity scores weighting using the boosted regression tree approach (Generalized Boosted Model).[9] This approach of estimating generalized propensity scores is similar to the one described above, but it was adjusted to accommodate multiple treatments. In the words of McCaffrey et al. (2013), the "proposed method estimates multiple causal effects by applying the binary tools multiple times. The balance metrics for ATE need to be adjusted to compare estimates with the pooled sample rather than to the other treatment mean" (p. 3412). In general, the employed approach is based on prior studies that extended propensity score methods to multiple treatments (Imbens, 2000; Imai & van Dyk, 2004; McCaffrey et al., 2013; Feng et al., 2012; Guo & Fraser, 2015; Cefalu & Buenaventura, 2017).

Event History Analysis

To address the above-described issues of time-dependency of data and censoring and properly operationalize the progression to college outcomes, we employed the method known as Event History Analysis (EHA aka duration analysis or survival analysis). Specifically, we used the Cox proportional

[9] For multiple treatment estimates, we used the *twang* package developed by the RAND corporation, which can handle multiple treatments via the *mnps* function (Cefalu & Buenaventura, 2017; McCaffrey et al., 2013).

hazards model to model time to graduation and time to university transfer. The dependent variable in EHA a hazard rate (or hazard), which is an instantaneous probability of experiencing the event of interest (earning a college credential or transferring to university) given that a student has not experienced this event until a given moment in time. In other words, a hazard is a rate at which events happen. In the *Results* section, we report hazard ratios, which are proportions of the hazards of the treated and untreated groups. The Cox model is based on the assumption of hazards being proportional over time. We checked for violation of this assumption and addressed it for problematic covariates by using time-varying covariates[10] in each EHA model specification. (Allison, 1984; Bennett, 1999; Box-Steffensmeier & Jones, 2004; Cleves et al., 2016).[11]

To summarize, this section discussed all main and secondary methodological approaches that were used in the study. The final models, which were chosen for analysis and whose findings are reported in the subsequent *Results* section, include the following components: generalized boosted modeling to estimate propensity scores, inverse probability of treatment weighting applied to all regression and EHA models (with normalized and truncated weights), average treatment effect as the counterfactual estimand, and dosage analysis for nonbinary treatments. Various comparison models used in the sensitivity analysis check whether the outcomes will be robust across different methods. Depending on the outcome variable and the model, we used the following control variables in doubly robust estimation: age (in the first term in logistic and OLS regression and time-varying in EHA), gender, dummy variables for race/ethnicity groups (Asian, Black, Hispanic, White, Other), Pell-ever status, ACT composite score, and dummy variable for college of enrollment. To account for the fact that weights were estimated and approximate the standard errors of the ATE and ATT estimates, we used robust sandwich standard errors and bootstrapping in some sensitivity analysis models (Thoemmes1 & Ong, 2015).

[10] Time-varying covariates are listed in the Notes of the results tables.

[11] EHA, in general, and the Cox model, in particular, are discussed in more details, including advantages over other methods, in the TBR's (2020b) working paper *Student engagement and college outcomes: Analysis of CCSSE data on Tennessee community college students* (https://www.tbr.edu/policy-strategy/presentations-and-papers)

Dataset description

This study relies on several data sources: student information system managed by TBR, which includes enrollment and graduation data for community college students and data on participation in high impact practices; student-level data from the National Student Clearinghouse (NSC), which allows tracking TBR students across institutional sectors and states; and Tennessee Student Assistance Corporation (TSAC),[12] which provided data on students who were eligible for a Pell grant at any time in their education history and students who received assistance via the Tennessee Promise program. The observation period for Iteration 2 of the study covers 12 calendar semesters: from fall 2017 through summer 2021.[13] The data on participation in service learning HIP was available for 11 terms, through spring 2021.

We compiled several datasets to address our research questions and use the respective models. All datasets include students from the cohort but may have different structures and contain additional variables. The cohort under analysis includes all first-time freshmen enrolling in TBR community colleges in fall 2017. It includes both full- and part-time enrolled students and students who enrolled as freshmen in summer 2017 and returned to college in fall. The cohort was also unduplicated by student ID: in case of simultaneous enrollment in two colleges in the first semester, the college with the highest number of attempted credits in that term was selected.

It is important to distinguish between the *full sample,* which is used in most descriptive analyses, and *analytic samples,* which are used in quantitative modeling. Both types of samples include the same number of freshmen, their educational history, and outcomes; however, they differ in terms of coding service learning participation. For descriptive analysis of the full sample, all service learning experiences are counted and reported, including the cases when students participated in multiple duration levels of the HIP (i.e., *Service Learning 1, 2,* and *3,* which are based on the number of hours in the component). In contrast, the subsequent *Results* section will report findings from the models that use a different approach: for each

[12] The researchers are grateful to THEC/TSAC leadership for making these data available.
[13] The two iterations of the study are described in the previous section *Secondary Models in Sensitivity Analysis.* Iteration 1 covered nine calendar semesters: from fall 2017 through summer 2020.

duration level, only participants in that level are included and there is a separate category (*Multiple Service Learning*) for students who participated in different duration levels. As a result, the number of service learning participants in analytic samples is smaller than their count in the full one. Also, students may participate in each duration level more than once, and the count of students in each frequency category differs for the full and analytic samples. **Table 1** presents the number of students in the full and analytic samples for all service learning participants and each of the duration level and frequency category.

Table 1. Service learning participation by duration level and frequency: Full and analytic samples *

	Full sample **				Analytic samples			
HIP participation	Any frequency	Once	Twice	3+ times	Any frequency	Once	Twice	3+ times
Any Service Learning	5,057	2,661	1,923	473	4,979	2,661	1,923	473
Service Learning 1 (< 10 hours)	2,970	2,446	391	133	1,309	985	262	95
Service Learning 2 (10 - 19 hours)	3,490	2,946	442	102	1,874	1,533	*353* **	
Service Learning 3 (20+ hours)	263	208	44	11	161	-	-	-
Multiple Service Learning	-	-	-	-	1,635	-	-	-
Nonparticipants	16,521							
Cohort	21,578							

* The data on HIP participation was available through spring 2021.
** Duplication across duration levels is possible in the full sample due to some students participating in different levels over time.
*** The categories "*Twice*" and "*3+ times*" for *Service Learning – 2* were combined into "*2+ times*" in the respective analytic sample due to small size of the group "*3+ times*".

There are 21,578 students in the first-time freshmen cohort, 5,057 (23.4 percent) of whom participated in service learning. Among participants in any service learning experience, 2,661 students (52.6 percent of participants) experienced it once, 1,923 students (38 percent) participated in service learning twice, and 473 students (9.3 percent of participants) took part in this HIP three or more times during the observation period. As explained above, some students experienced service learning of different durations. Because of this, in the full sample, there exists some duplication on student ID across *Service*

Learning 1, *2*, and *3*, and the sum of students in each duration level exceeds the total number of participants. However, this duplication is addressed in the analytic samples: all students who experienced HIP of different durations are placed under *Multiple Service Learning*, which is analyzed as a separate group of participants. In both full and analytic samples, the majority of students participated in *Service Learning 2*, which is a course or section with a service learning component of 10 to 19 hours of service. The smallest duration level group is *Service Learning 3* (20 or more hours): 263 students in the full sample and 173 students in the analytic sample. Because breaking down this category by frequency of participation leads to very small groups, this duration level is not used in models that require frequencies and is omitted for analytic samples in Table 1. To reiterate, analytic samples are used in quantitative analyses that examine the effect of service learning participation on the outcomes of focus. It is important to keep sample size in mind when interpreting the results of the descriptive analysis by duration level.

Table 2 presents demographic and academic variables for the entire cohort and by service learning participation and identifies the ones with statistically significant difference between HIP participants and nonparticipants. These comparisons show that service learning participants differ from nonparticipants on a number of characteristics and there are differences by duration level. For racial/ethnic groups, there is a large and statistically significant difference between service learning participation of Black and white students. Namely, there are fewer Black students and more white students among participants (8.9 and 79 percent, respectively) than among nonparticipants (17.9 and 70 percent, respectively). This difference is the largest for *Service Learning 3* (4.6 and 81 percent, respectively), but it is also observable for the other duration levels. The share of female service learning participants (57.5 percent) is higher than the share of male participants (42.5 percent); it is the highest for students taking *Service Learning 3* (66.2 percent). Service learning participants have fewer adult students (5.2 versus 8.3 percent) and Pell-eligible students (60.2 versus 65.8 percent) as compared to nonparticipants. By duration level, the share of adult students is the smallest for *Service Learning 2* participants (4.8 percent), and the share of Pell-eligible students is the smallest among *Service Learning 3* participants (51.7 percent).

Table 2. Participation in service learning by demographic and academic variables

		Any SL		SL 1		SL 2		SL 3		Nonparticipants	
			5,057		2,970		3,490		263		16,599
	Cohort	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%
Asian	314	75	1.5	49	1.7	55	1.6	6	2.3	239	1.5
Black #	3,872	448	8.9	340	11.5	280	8.0	12	4.6	3,424	20.7
Hispanic	1,339	322	6.4	220	7.4	236	6.8	15	5.7	1,017	6.2
White #	15,093	3,996	79.0	2,247	75.7	2,758	79.0	213	81.0	11,097	67.2
Other	960	216	4.3	114	3.8	161	4.6	17	6.5	744	4.5
Male #	9,475	2,152	42.5	1,259	42.4	1,487	42.6	89	33.8	7,323	44.3
Female #	12,103	2,905	57.5	1,711	57.6	2,003	57.4	174	66.2	9,198	55.7
Tradit'l age #	19,941	4,794	94.8	2,823	95.1	3,321	95.2	245	93.2	15,147	91.7
Adult #	1,637	263	5.2	147	4.9	169	4.8	18	6.8	1,374	8.3
Non-Pell ever #	7,657	2,011	39.8	1,133	38.2	1,374	39.4	127	48.3	5,646	34.2
Pell ever #	13,921	3,046	60.2	1,837	61.8	2,116	60.6	136	51.7	10,875	65.8
Non-learning support #	7,749	1,680	33.2	796	26.8	1,061	30.4	127	48.3	6,069	36.7
Learning support #	13,829	3,377	66.8	2,174	73.2	2,429	69.6	136	51.7	10,452	63.3
Non-Promise #	7,952	1,405	27.8	853	28.7	924	26.5	61	23.2	6,547	39.6
Promise #	13,626	3,652	72.2	2,117	71.3	2,566	73.5	202	76.8	9,974	60.4
		Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median
High school GPA #		3.12	3.12	3.09	3.08	3.10	3.10	3.28	3.43	2.99	2.96
ACT score #		19.5	19.0	18.9	19.0	19.3	19.0	21.0	21.0	19.0	19.0
Final GPA #		2.51	2.62	2.41	2.49	2.48	2.56	3.11	3.29	2.22	2.05
Average credits earned #		42.6	47.0	40.1	42.0	41.2	44.0	63.8	64.0	31.3	24.0

Note. Percent is of the total for each demographic breakdown and service learning type.
# Indicates variables with a statistically significant difference between service learning participants (any service learning)and nonparticipants (*p*<0.001) based on *chi-square* tests or two-independent samples *t*-tests for difference in means (two-sided).

Table 2 also shows that there is a statistically significant difference between service learning participants and nonparticipants on key academic variables. On average, HIP participants have higher high school GPA that nonparticipants (3.12 versus 2.99), larger ACT composite score (19.5 versus 19.0), higher cumulative college GPA in the last semester of observation (2.5 versus 2.2), and more average cumulative credits earned (42.6 versus 31.3). Again, at a purely descriptive level, students taking part in the longer *Service Learning 3* duration level tend to have higher high school and college GPA and ACT score than other participants; on average, they also accumulate many more cumulative credits than participants in other duration levels and nonparticipants. Service learning participants also include more learning support students than their counterparts who did not partake of service learning opportunities (66.8 versus 63.3 percent) and more Tennessee Promise students (72.2 versus 60.4 percent). The share of students who were in need of learning support is the highest among *Service Learning 1* participants (73.2 percent). The *Service Learning 3* participants have the largest share of Promise students (76.8 percent), followed by students in *Service Learning 2* (73.5 percent).

Table 3 presents participation in service learning in general and by duration level and semester of enrollment. It shows that most frequently students who start out as first-time freshmen take part in service learning in their first semester: 52.7 percent of all cases during the observation period for any duration, 56.8 percent of all cases for *Service Learning 1*, and 64.3 percent for *Service Learning 2*. It should be noted that in some TBR community colleges service learning is a component of, or is tied to, the first-year experience. The second semester of enrollment (spring 2018) follows with 16.0, 16.3, and 12.3 percent of all cases for these duration levels, respectively. In contrast, for *Service Learning 3*, participation is more evenly spread over time, with calendar semester 7 (fall 2019) marking the peak of participation with 19.8 percent of all cases of participation during the observation period. Importantly, the group of nonparticipants in Table 3 includes ***all*** students who did not participate in service learning only for the *Any Service Learning* category; however, it is ***level-specific*** for duration levels *Service Learning 1*, *2*, and *3*. In other words, for a particular duration level, nonparticipants include students who did not participate in that level specifically over the entire period of observation.

Table 3. Participation in service learning by duration level and semester

		Fall 2017	**Spring 2018**	**Summer 2018**	**Fall 2018**	**Spring 2019**	**Summer 2019**	**Fall 2019**	**Spring 2020**	**Summer 2020**	**Fall 2020**	**Spring 2021**
Any Service Learning												
Participants	N	3,381	1,023	33	636	699	26	274	177	22	75	68
	%	**52.7**	**16.0**	0.5	9.9	10.9	0.4	4.3	2.8	0.3	1.2	1.1
Nonparticipants	N	7,901	4,801	382	2,834	2,158	370	1,189	783	109	394	351
	%	37.1	22.6	1.8	13.3	10.1	1.7	5.6	3.7	0.5	1.9	1.7
Service Learning 1												
Participants	N	1,964	564	3	340	294	21	115	73	10	34	38
	%	**56.8**	**16.3**	0.1	9.8	8.5	0.6	3.3	2.1	0.3	1.0	1.1
Nonparticipants	N	9,318	5,260	412	3,130	2,563	375	1,348	887	121	435	381
	%	38.5	21.7	1.7	12.9	10.6	1.6	5.6	3.7	0.5	1.8	1.6
Service Learning 2												
Participants	N	2,587	497	32	283	384	5	117	72	5	30	14
	%	**64.3**	**12.3**	0.8	7.0	9.5	0.1	2.9	1.8	0.1	0.8	0.4
Nonparticipants	N	8,695	5,327	383	3,187	2,473	391	1,346	888	126	439	405
	%	36.8	22.5	1.6	13.5	10.5	1.7	5.7	3.8	0.5	1.9	1.7
Service Learning 3												
Participants	N	47	39	1	41	48	2	63	37	8	13	20
	%	14.7	12.2	0.3	12.9	**15.1**	0.6	**19.8**	11.6	2.5	4.1	6.3
Nonparticipants	N	11,235	5,785	414	3,429	2,809	394	1,400	923	123	456	399
	%	41.1	21.1	1.5	12.5	10.3	1.4	5.1	3.4	0.5	1.7	1.5

Note. Percentage is of all terms of observation. Duplication on ID is possible due to students participating in multiple service learning options in different semesters.

Table 4 provides service learning participation by major in the first term of this HIP experience. Overall, when students participate in any service learning, they tend to be enrolled in the following major fields: *Liberal Arts and Sciences, General Studies and Humanities* (72.7 percent); *Education, General* (5 percent); *Business Administration, Management and Operations* (4.7 percent); *Registered Nursing, Nursing Administration, Nursing Research and Clinical Nursing* (2.9 percent); *Computer and Information Sciences, General* (2 percent), and *Human Development, Family Studies, and Related Services* (1.6 percent).

There is some variation in major-taking patterns among duration levels, most notably for *Service Learning 3*. For this group, there is a smaller share of students who enrolled in *Liberal Arts and Sciences, General Studies and Humanities* in the term when they participated in this HIP: 43.3 percent as compared to 72.7 percent for any service learning participants. In addition to Liberal Arts and Sciences, students completing *Service Learning 3* were also often enrolled in *Allied Health Diagnostic, Intervention, and Treatment Professions* (14.1 percent); *Business Administration, Management and Operations* (12.2 percent); *Communications Technology/ Technician* (6.9 percent); and *Visual and Performing Arts* (4.1 percent). In contrast, participants in *Service Learning 1* and *2* in the cohort under analysis seldom enrolled in the last three fields. To reiterate, Table 4 presents majors at the time of the first service learning experience and not necessarily the program in which a student started out in the first semester.

The remainder of this section will examine various outcomes for the entire cohort and by service learning participation and duration level. It is worth reminding here that while the observation period for the study covers twelve calendar semesters (fall 2017 through summer 2021), the data on service learning participation was available for eleven semesters—through spring 2021. However, the data for all reported outcomes—both for the TBR and the National Student Clearinghouse data—include twelve terms of observation.

Figure 1 demonstrates key college outcomes for the 2017 first-time freshmen cohort tracked for twelve calendar semesters, from fall 2017 through summer 2021.

Table 4. Major in the first term of service learning participation

Major	Any SL		SL 1		SL 2		SL 3	
	N	%	N	%	N	%	N	%
Liberal Arts & Sciences, General Studies	4,655	72.69	2,668	77.31	3,011	74.88	138	43.26
Education, General	319	4.98	95	2.75	289	7.19	1	0.31
Business Administration, Management	303	4.73	152	4.40	146	3.63	39	12.23
Registered Nursing, Nursing Administration	184	2.87	99	2.87	102	2.54	6	1.88
Computer & Information Sciences, General	127	1.98	53	1.54	80	1.99	1	0.31
Human Development, Family Studies	104	1.62	56	1.62	43	1.07	12	3.76
Allied Health Diagnostic, Intervention	86	1.34	12	0.35	32	0.80	45	14.11
Industrial Production Technologies	85	1.33	31	0.90	60	1.49	1	0.31
Criminal Justice & Corrections	80	1.25	31	0.90	65	1.62		
Music	73	1.14	61	1.77	18	0.45	2	0.63
Communications Technology / Technician	46	0.72	4	0.12	22	0.55	22	6.90
Electromechanical Instrumentation	39	0.61	38	1.10	32	0.80		
Electrical Engineering Technologies	37	0.58	12	0.35	26	0.65		
Allied Health & Medical Assisting Services	31	0.48	18	0.52	3	0.07	10	3.13
Health Professions & Related Clinical Science	31	0.48	23	0.67	7	0.17	3	0.94
Engineering Technology, General	28	0.44	19	0.55	10	0.25		
Dental Support Services & Allied Professions	25	0.39	15	0.43	6	0.15	7	2.19
Business Operations Support	22	0.34	12	0.35	11	0.27		
Audiovisual Communications Technologies	21	0.33	2	0.06	21	0.52		
Clinical/Medical Laboratory Science/Research	16	0.25	12	0.35	13	0.32		
Visual & Performing Arts, General	14	0.22	1	0.03			13	4.08
Vehicle Maintenance & Repair Technologies	13	0.20	3	0.09	11	0.27		
Hospitality Administration / Management	12	0.19	12	0.35				
Public Administration and Social Service	10	0.16	10	0.29				
Agricultural & Domestic Animal Services	9	0.14					9	2.82
Design and Applied Arts	7	0.11	4	0.12	2	0.05	1	0.31
Multi-/Interdisciplinary Studies, General	6	0.09	2	0.06	5	0.12		
Health & Medical Administrative Services	4	0.06					4	1.25
Legal Support Services	4	0.06	2	0.06	2	0.05		
Ophthalmic & Optometric Support Services	4	0.06					4	1.25
Drafting / Design Engineering Technologies	3	0.05			3	0.07		
Architectural Engineering Technologies	1	0.02	1	0.03				
Basic Skills & Developmental Education	1	0.02	1	0.03				
Engineering, General	1	0.02					1	0.31
Health Services / Allied Health	1	0.02	1	0.03				
HVACR Maintenance Technology	1	0.02			1	0.02		
Mechanical Engineering Related Technologies	1	0.02	1	0.03				

Note. Duplication on ID is possible due to some students participating in different duration levels of service learning over time.

Figure 1. Outcomes for the 2017 first-time freshmen cohort

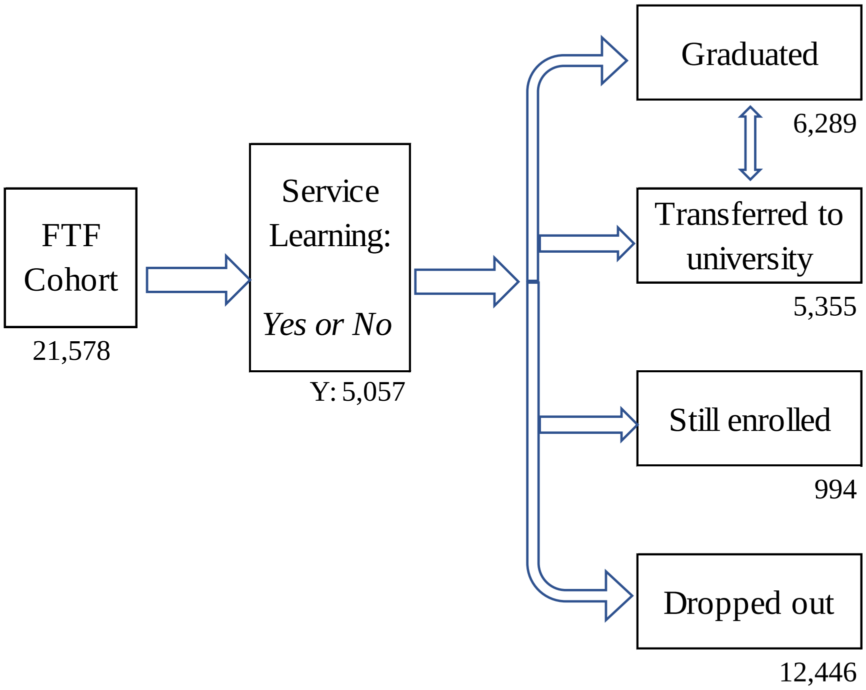


Figure 1 shows that out of 21,578 students in the cohort, 5,057 (23.4 percent) took part in service learning. The outcomes are presented for the entire cohort regardless of service learning experience. By summer 2021, 6,289 students (29.1 percent) earned a college credential and 5,355 students (24.8 percent) transferred to a four-year college or university.[14] It is important that these outcomes are not mutually exclusive, and 3,506 students (16.2 percent) both graduated and transferred to university; this dynamic is shown by the reverse arrow. Graduation, which is understood as earning a technical certificate, associate degree, or bachelor's degree, may precede or follow transfer. In spring or summer of 2021, 994 students (4.6 percent) were still enrolled in TBR community colleges.[15] Finally, 12,446 students (57.7 percent of the cohort) either dropped out or stopped out and were no longer available for observation.

[14] Either in Tennessee or other states (limited to higher education institutions submitting data to the National Student Clearinghouse).

[15] Iteration 1 of the study used a more stringent definition of *Still Enrolled* students, which also relied on a minimum number of terms of enrollment in the TBR system. As a result, the counts of *Still Enrolled* and *Dropped Out* students are not directly comparable across the study iterations.

Figure 2 shows general outcomes by service learning participation. **Table 5** presents the main outcomes of interest, graduation and transfer, in more details and by service learning duration level.

Figure 2. Outcomes by service learning participation

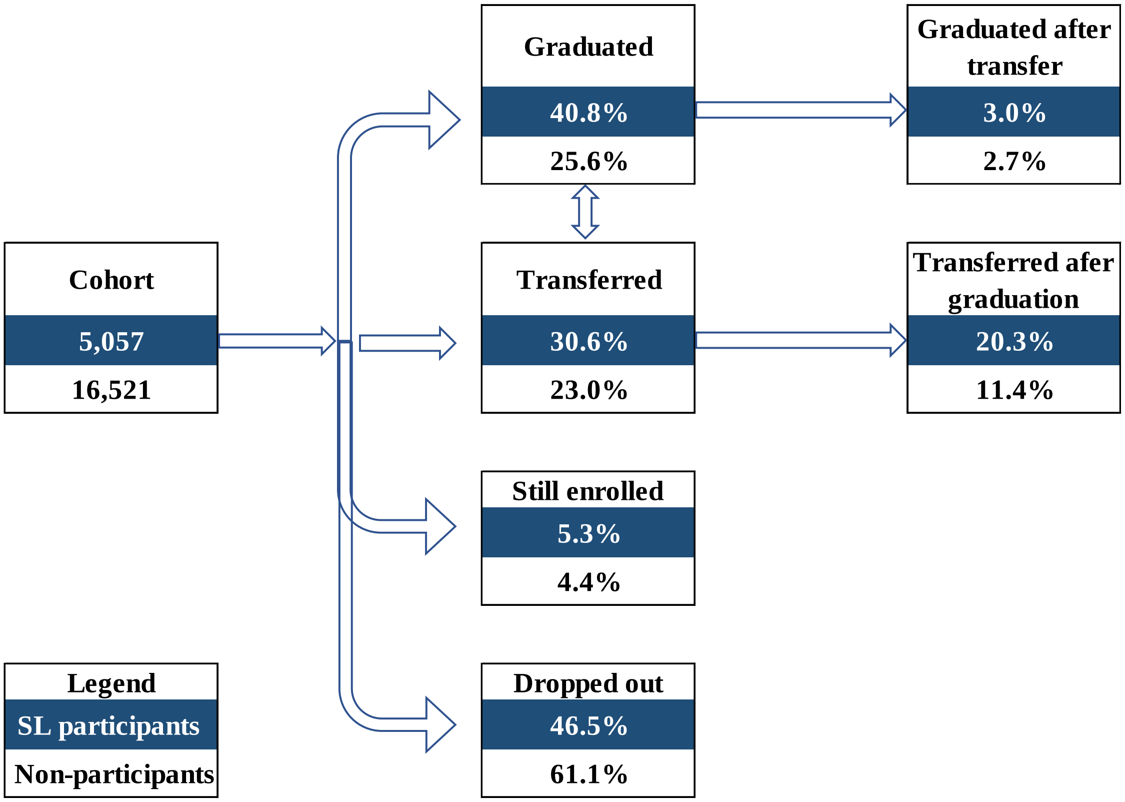


Figure 2 and Table 5 demonstrate that at a purely descriptive level, service learning participants and nonparticipants differ in how they attain outcomes. Larger shares of participants than nonparticipants graduate (40.8 vs. 25.6 percent) and transfer to university (30.6 vs. 23 percent), or are still enrolled at the end of the observation period (5.3 vs. 4.4 percent). As a result, a smaller share of participants is among students who dropped out or stopped out: 46.5 percent as compared to 61.1 percent for nonparticipants. It is noteworthy that a much larger share of service learning participants transfers to university after earning a credential at a community college: 20.3 percent of participants transfer after graduation as opposed to 11.4 percent of nonparticipants who do the same. This observation is related to the results of some quantitative analyses, which are presented in the subsequent *Results* section.

Table 5. Outcomes by service learning participation and duration level

		Graduated		Transferred		Both outcomes	
	Total	All graduates	Did not transfer	All Transfers	Did not graduate	Graduation before transfer	Transfer before graduation
Any SL	5,057	2,062	887	1,549	374	1,024	151
		40.8%	17.5%	30.6%	7.4%	20.3%	3.0%
SL 1	2,970	1,087	475	829	217	528	84
		36.6%	16.0%	27.9%	7.3%	17.8%	2.8%
SL 2	3,490	1,356	542	1,065	251	705	109
		38.9%	15.5%	30.5%	7.2%	20.2%	3.1%
SL 3	263	204	114	100	10	88	2
		77.6%	43.4%	38.0%	3.8%	33.5%	0.8%
Not any SL	16,521	4,227	1,896	3,806	1,475	1,888	443
		25.6%	11.5%	23.0%	8.9%	11.4%	2.7%

Note. Percentage of the group's total. Duplication on ID is possible due to participation in multiple service learning levels.

Examination by duration level demonstrates that *Service Learning 3* participants have the highest rates of success as measured by share of graduates (77.6 percent), students who transfer to a four-year college or university (38 percent), and graduates who earned a credential prior to transferring to university (33.5 percent). While having lower percentage for these outcomes, participants in *Service Learning 1* and *2* outperform nonparticipants on all of them. For these two duration levels, the share of students who graduate before university transfer also exceeds the one for nonparticipants: 17.8 percent for *Service Learning 1* and 20.2 percent for *Service Learning 2* as opposed to 11.4 percent for nonparticipants.

Time to an outcome can be measured in either semesters or credits. **Table 6** provides average time to graduation (any college credential versus associate degrees or higher) and university transfer as measured in semesters of enrollment and cumulative credits attempted by service learning duration levels and key demographic and academic variables. For baseline comparison, it also shows terms-to-outcome and credits-to-outcome for all graduates and transfer students in the cohort. **Figure 3** presents time to graduation by demographic variables.

For graduation, time to graduation is similar across main comparison groups when measured in semesters of enrollment. On average, all graduates, service learning participants and nonparticipants take about five and a half semesters before they earn a college credential (either a technical certificate or a degree), with participants graduating slightly faster. The difference between participants and nonparticipants becomes a little more pronounced when time to a college degree only is considered but remains small: on average, service learning students earn a degree in 5.6 terms and nonparticipants do so in about 5.8 semesters. When measuring time to graduation in attempted credits, service learning participants attempt more credit hours (72.2 credits), on average, than nonparticipants (70.4 credits) before graduating.

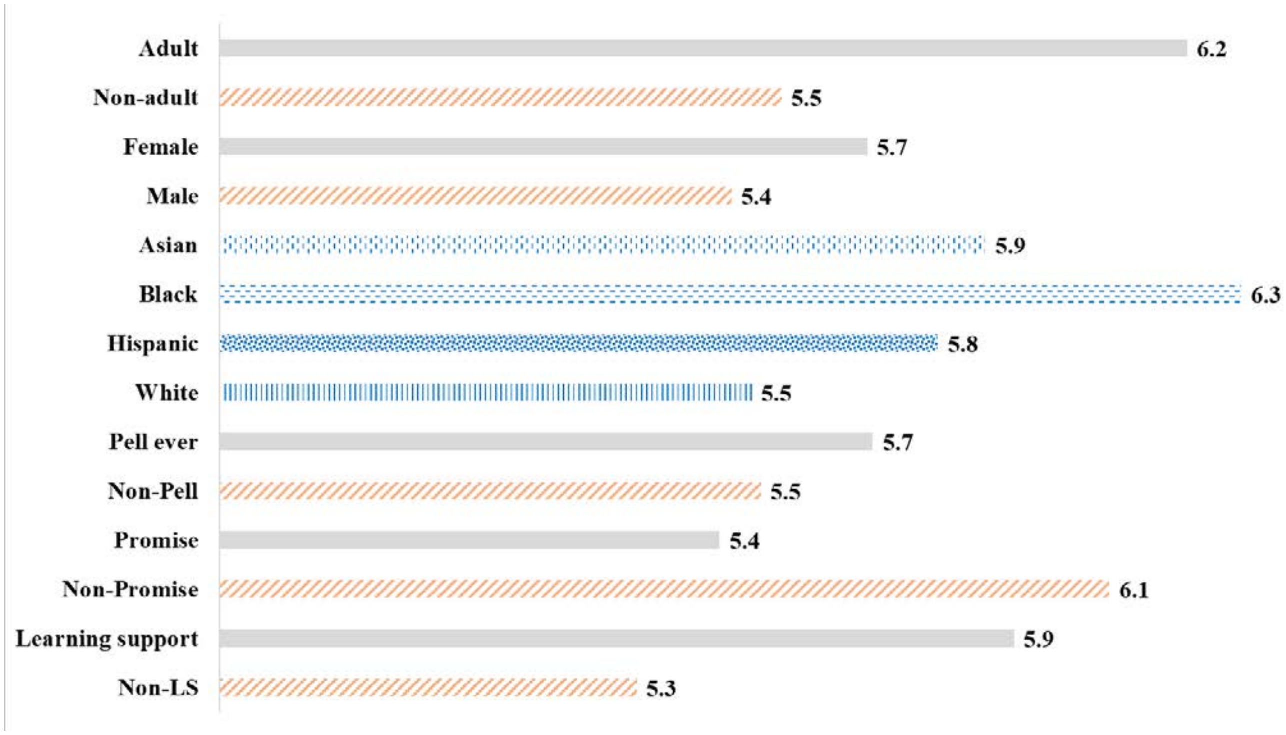
There is also some variability by duration level. *Service Learning 3* completers demonstrate the shortest mean time to graduation (5.1 terms to any award and 5.3 terms to a degree) but also the largest number of attempted credits (75 credits to any award and 76.4 credits to a degree). In contrast, students who participated in multiple durations of service learning have the longest mean time to graduation in semesters (5.7 terms to any award and 5.8 terms to a degree) and the second largest number of attempted credits (72.7 credits to any award and 72.5 credits to a degree).

For transfer, nonparticipants, though, demonstrate the shortest average time (5.3 terms) and the smallest number of attempted credits (57.9 credit hours) to this outcome. Overall service learning participants and participants by duration level have longer mean time to university transfer—measured in either semesters of enrollment or attempted credits—than nonparticipants or all transfer students.

Table 6. Average time to graduation and university transfer in terms and credits attempted

	Mean semesters and attempted credits to …					
	Graduation				Transfer	
	Any award (5,163)		Degrees (4,497)		(4,159)	
	Terms	Attempted credits	Terms	Attempted credits	Terms	Attempted credits
Total by outcome	5.57	71.0	5.71	71.3	5.40	60.0
Service learning (SL) levels						
Any SL	5.50	72.2	5.63	72.2	5.74	65.0
SL - 1	5.64	71.9	5.76	71.8	5.69	63.8
SL - 2	5.51	72.2	5.61	72.1	5.76	65.2
SL - 3	5.12	75.0	5.33	76.4	6.05	71.8
Multiple SL	5.67	72.7	5.77	72.5	5.76	64.5
Not any SL	5.60	70.4	5.75	70.9	5.26	57.9
Demographic and academic variables						
Adult in Term 1	6.18	74.7	6.68	77.5	5.95	58.3
Trad. age in Term 1	5.52	70.9	5.65	71.0	5.38	60.0
Female	5.66	71.6	5.75	71.8	5.46	60.2
Male	5.44	70.2	5.65	70.5	5.33	59.6
Asian	5.85	72.4	5.91	71.2	5.97	64.8
Black	6.27	74.6	6.41	75.1	5.18	53.0
Hispanic	5.77	72.1	5.98	73.2	5.48	60.5
White	5.47	70.4	5.60	70.7	5.43	61.2
Other race	5.65	73.2	5.76	73.3	5.18	56.0
Pell ever	5.67	72.0	5.83	72.5	5.39	58.8
Non-Pell	5.48	70.1	5.60	70.2	5.41	60.9
Promise	5.42	71.3	5.53	71.4	5.54	63.1
Non-Promise	6.05	70.1	6.32	71.0	5.05	51.9
Learning support	5.90	72.1	6.12	73.0	5.46	57.8
Non-learning support	5.28	70.0	5.38	70.0	5.35	61.9

Figure 3. Time to graduation by demographic variables



In addition to time to outcome by service learning participation and nonparticipation, Table 6 and Figure 3 also demonstrate how time to outcome differs by demographic and academic categories. The following student groups have a shorter average time to graduation than their counterparts: traditional-age, male and white students, individuals who were not Pell-eligible at any time, Tennessee Promise students, and students who did not require learning support. The differences by demographic and academic variables are also observed for attempted credits to graduation as well as for university transfer as an outcome.

These differences, as well as the other results of the descriptive analysis in this section, clearly demonstrate the need to account for demographic and academic factors in quantitative models that aim to estimate the effect of service learning participation on educational outcomes of interest. The preceding *Methodology* section has explained the empirical strategies that are employed to address the research questions of the study while accounting for differences among various categories of students, and service learning participants versus nonparticipants.

Results

The *Methodology* section describes different approaches to estimating the effect of service learning participation. Based on the nature of treatment variables, the employed strategies for the main models fall into two main categories. For binary treatments, the inverse probability of treatment weighting was used; for non-binary treatment, dosage analysis with generalized propensity scores was employed. The subsequent discussion of findings proceeds by outcome rather than by the broad method, with dichotomous and multiple treatment conditions presented in turn for the same outcome of interest. We provide outcomes for the unweighted and weighed samples for comparison but discuss statistically significant findings for the weighted samples only. Besides, results for models with and without control variables are presented; however, the discussion is limited to findings from doubly robust models that used control variables and truncated inverse probability of treatment weights. Appendices 5-9 offer findings from the secondary models that were used in the specification analysis (as explained in *Methodology*, specification analyses were run in Iteration 1 of the study, which covered nine terms of tracking data).

This section starts with logistic and OLS regression models for the following outcomes: the probability of graduation, the probability of transfer to a four-year college or university, the probability of student departure (dropping out or stopping out as of the last term), and final GPA. The following control variables were used in these models: age in the first semester, dummy variable for gender, race/ethnicity indicator variables, Pell-ever status flag, ACT composite score, and college in the first term. The Event History Analysis (EHA) models, which are presented afterwards, estimate hazards for graduation, university transfer, and student departure and use the following control variables: age in each term, dummy variable for gender, race/ethnicity indicator variables, Pell-ever status flag, ACT composite score, and college of enrollment. Depending on the EHA model, outcome, and results of tests for proportional hazard assumption violation, some control variables were modified as time-varying covariates to include time-dependent effect. All models estimate average treatment effect using the inverse probability of treatment weights, which were truncated at 1% and 99% to attenuate the effect of large weights.

The first set of models estimates the probability of graduation due to service learning participation. It is important to keep in mind—for all tables of results in this section—that different weighting and modeling approaches were used for each group of treatment variables, binary and non-binary.

Table 7 presents predicted increase in the probability of graduation as a result of service learning participation both for binary and non-binary treatments. We find that taking part in any service learning opportunity increases the predicted probability of earning a college credential (technical certificate or degree) by 17 percentage points. This general effect differs by duration level: the likelihood of graduation is predicted to increase by 16 percentage points for *Service Learning 1*, 15 percentage points for *Service Learning 2*, 33 percentage points for *Service Learning 3*, and 17 percentage points for *Multiple Service Learning* durations experienced during the observation period. **Figure 4** shows the predicted impact of completing any service learning and various duration levels of this HIP on the probability of graduation together with the respective 95 percent confidence intervals.

Figure 4. Predicted increase in probability of graduation: Binary treatments

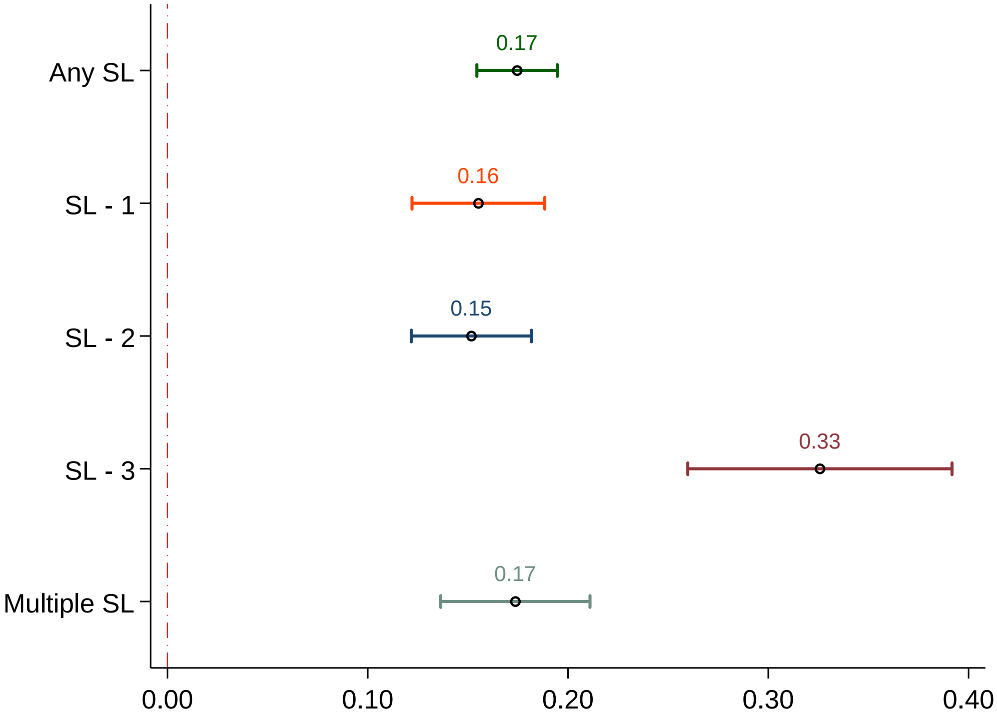


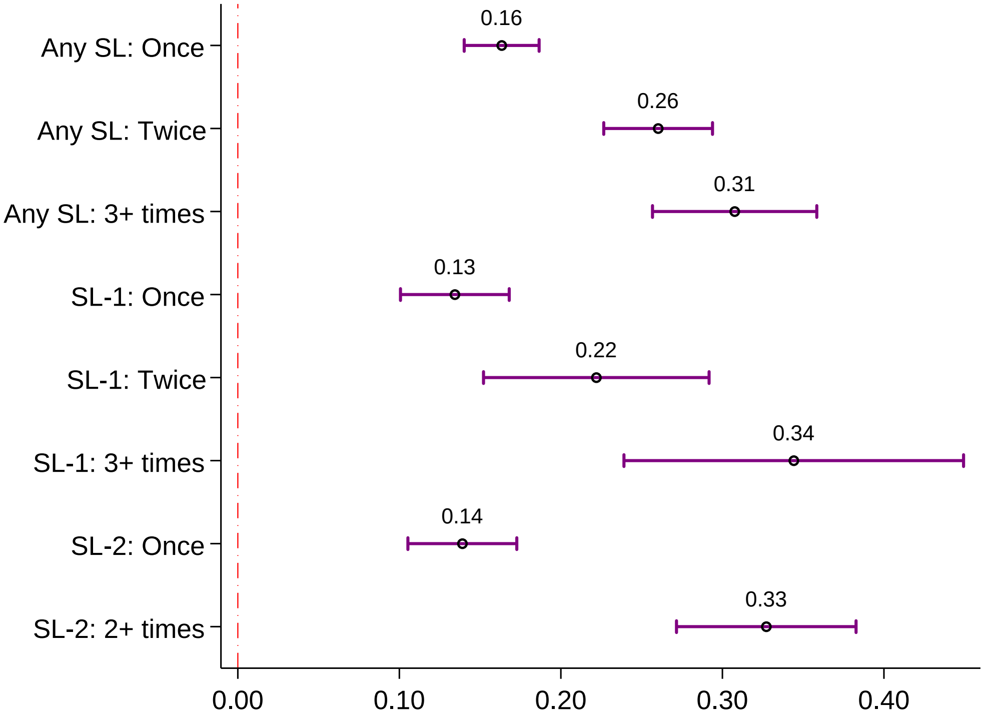
Table 7. Predicted increase in probability of graduation as a result of service learning participation

Service learning types, duration, and frequency		Unweighted Sample		Weighted Sample #	
		Margin	SE	Margin	SE
Binary treatment (Reference group: nonparticipants)					
Any Service Learning		0.14 ***	(0.01)	0.13 ***	(0.01)
Any Service Learning (with control variables)		0.15 ***	(0.01)	0.17 ***	(0.01)
Service Learning 1		0.10 ***	(0.01)	0.11 ***	(0.01)
Service Learning 1 (with control variables)		0.13 ***	(0.01)	0.16 ***	(0.02)
Service Learning 2		0.14 ***	(0.01)	0.12 ***	(0.01)
Service Learning 2 (with control variables)		0.10 ***	(0.01)	0.15 ***	(0.02)
Service Learning 3		0.40 ***	(0.04)	0.40 ***	(0.04)
Service Learning 3 (with control variables)		0.33 ***	(0.03)	0.33 ***	(0.03)
Multiple SL		0.06 ***	(0.01)	0.14 ***	(0.02)
Multiple SL (with control variables)		0.05 *	(0.01)	0.17 ***	(0.02)
Non-binary treatment (Reference group: 0 times)					
Any Service Learning:	Once	0.16 ***	(0.01)	0.12 ***	(0.01)
Any Service Learning:	Twice	0.11 ***	(0.01)	0.16 ***	(0.01)
Any Service Learning:	Three + times	0.27 ***	(0.02)	0.22 ***	(0.03)
Any Service Learning:	Once (with controls)	0.15 ***	(0.01)	0.16 ***	(0.01)
Any Service Learning:	Twice (with controls)	0.14 ***	(0.01)	0.26 ***	(0.02)
Any Service Learning:	Three + times (with controls)	0.28 ***	(0.02)	0.31 ***	(0.03)
Service Learning 1:	Once	0.12 ***	(0.02)	0.09 ***	(0.02)
Service Learning 1:	Twice	0.07 *	(0.03)	0.06 *	(0.03)
Service Learning 1:	Three + times	0.14 **	(0.05)	0.17 ***	(0.05)
Service Learning 1:	Once (with controls)	0.13 ***	(0.02)	0.13 ***	(0.02)
Service Learning 1:	Twice (with controls)	0.18 ***	(0.04)	0.22 ***	(0.04)
Service Learning 1:	Three + times (with controls)	0.25 ***	(0.05)	0.34 ***	(0.05)
Service Learning 2:	Once	0.12 ***	(0.01)	0.08 ***	(0.01)
Service Learning 2:	Two + times	0.35 ***	(0.03)	0.32 ***	(0.03)
Service Learning 2:	Once (with controls)	0.07 ***	(0.01)	0.14 ***	(0.02)
Service Learning 2:	Two + times (with controls)	0.24 ***	(0.03)	0.33 ***	(0.03)

* p < .05, ** p < .01, *** p < .001. Robust standard errors in parentheses.
Control variables: age in the first term, gender, race/ethnicity groups, Pell-ever status, ACT composite score, college in first term.
# Average Treatment Effect estimated with normalized Inverse Probability of Treatment Weights truncated at 1% and 99%.

The second pane of **Table 7** presents results for non-binary treatments, that is, frequency of service learning overall and by duration level. We find that service learning participation with different frequency is positively related to the likelihood of earning a credential. The probability of graduation increases by 16 percentage points if any service learning was experienced once, 26 if taken twice, and 31 percentage points if completed three or more times. For duration level 1, the likelihood of graduation is predicted to increase by 13 percentage points for participating once, 22 percentage points for taking part twice, and 34 percentage points for participating three or more times. For duration level 2, the completion likelihood increases by 14 percentage points when service learning was taken once and by 33 percentage points when service learning was taken two or more times.[16] We did not analyze *Service Learning 3* by frequency due to small sample size. **Figure 5** presents the impact of completing service learning on the probability of graduation by frequency. It also shows how confidence intervals depend on the analytic sample's size.

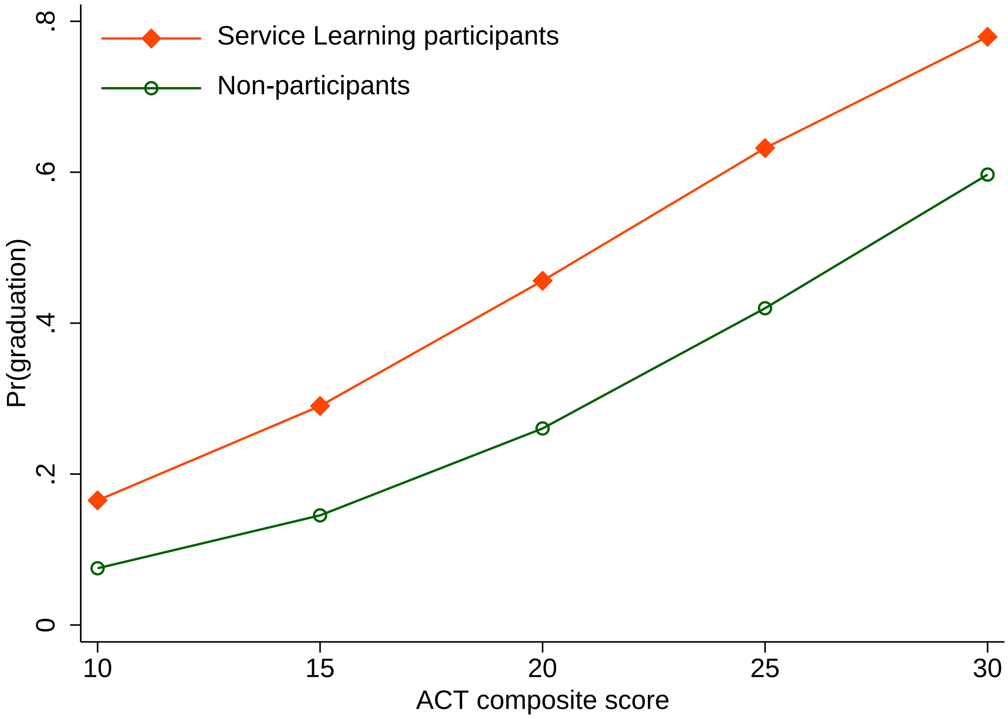
Figure 5. Predicted increase in probability of graduation: Non-binary treatments



[16] As a reminder, for *Service Learning – 2*, the categories "*Twice*" and "*3+ times*" were combined into "*2+ times*" due to small size of the latter group.

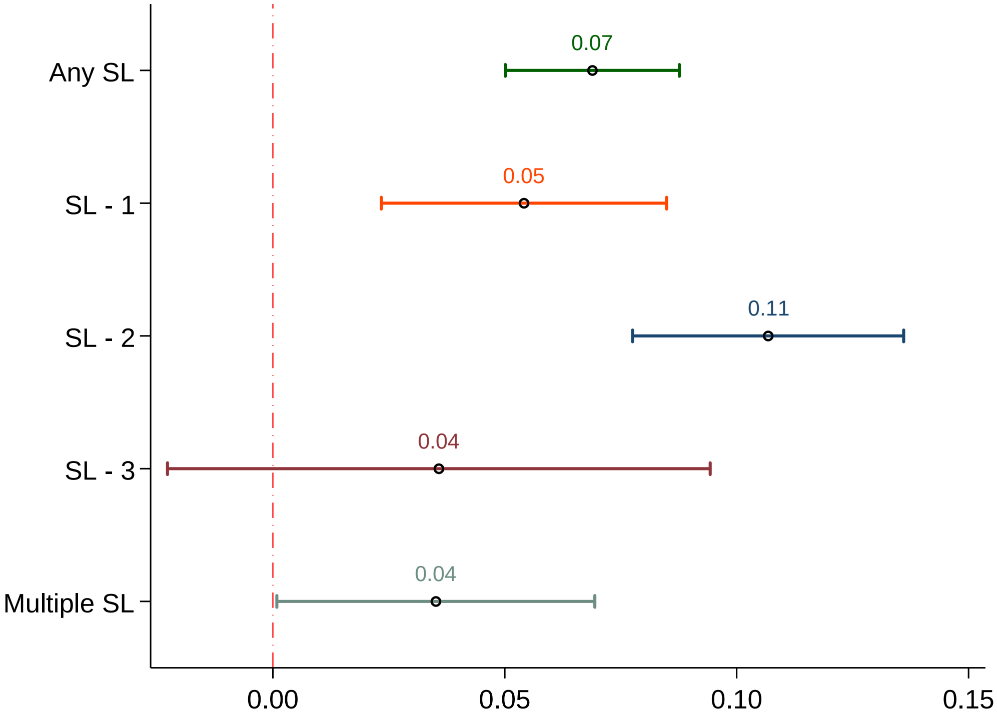
Figure 6 plots average marginal effects for service learning participants and nonparticipants by ACT composite score in the logistic model for graduation. Average marginal effects are predictions that are adjusted for actual observed values of covariates (not the mean values) in the model. In this approach to computing marginal effects, service learning participants and nonparticipants differ only in the treatment exposure but are otherwise identical "average" subjects on all the control variables. Of course, they are also similar on the inverse probability of treatment weights that are applied to the final model. Figure 6 shows that service learning participants have a higher predicted probability of graduation than nonparticipants for each ACT score category. It also shows that the gap between these student groups grows with an increase in the ACT composite score, except for the ACT score of 30 where the gap decreases slightly—likely due to a much smaller sample size.

Figure 6. Adjusted predictions for graduation for service learning participants and nonparticipants



The second set of models examined the impact of service learning participation on the probability of transfer to a four-year institution. **Table 8** and **Figures 7** and **8** present all findings for this outcome for the unweighted and weighted samples, and binary and non-binary treatments. Participation in any service learning is predicted to increase the probability of transfer by 7 percentage points. Completing duration levels 1 and 2 raises the likelihood of transfer by 5 and 11 percentage points, respectively. Multiple HIP experiences are predicted to increase the probability of transfer by 4 pp. The baseline model (without control variables) for *Service Learning 3* showed effects for the binary treatment, but it was no longer statistically significant in the doubly robust model with the final results. We find that the following frequency of service learning participation has a statistically significant positive effect on the likelihood of transfer (reported in Figure 8): *Any Service Learning* once (6 pp.), twice (9 pp.), and three or more times (9 pp.); *Service Learning 1* once (5 pp); and *Service Learning 2* once (9 pp) and two or more times (24 pp.).

Figure 7. Predicted increase in probability of transfer: Binary treatments



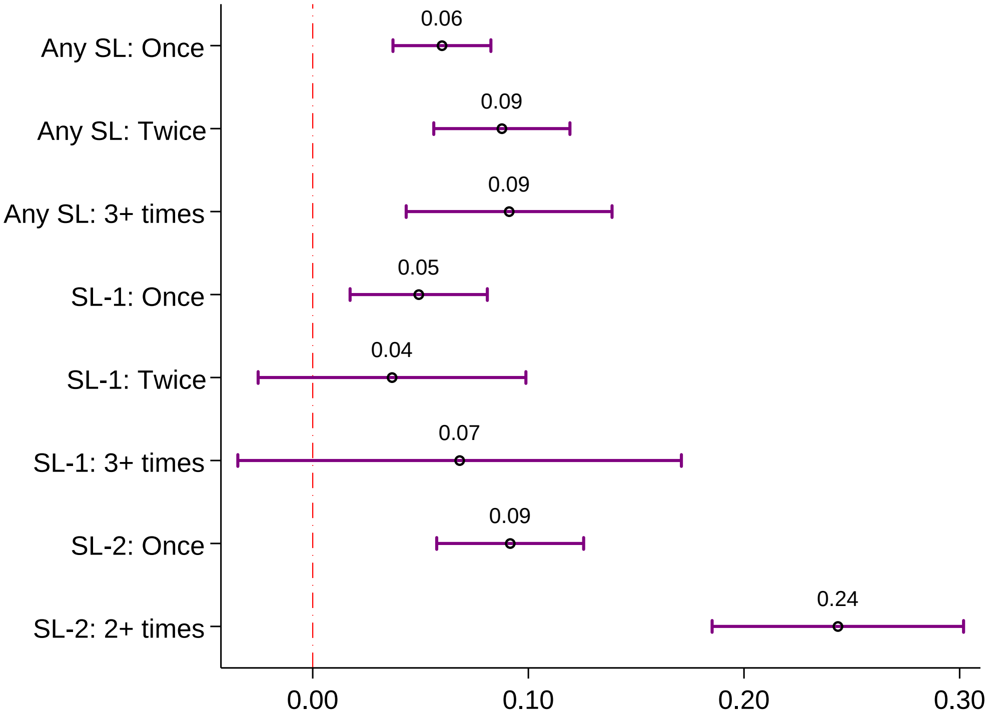
Note: If the 95% confidence interval crosses the vertical red line, the result is not statistically significant.

Table 8. Predicted increase in probability of transfer as a result of service learning participation

Service learning types, duration, and frequency		Unweighted Sample		Weighted Sample #	
		Margin	SE	Margin	SE
Binary treatment (Reference group: nonparticipants)					
Any Service Learning		0.07 ***	(0.01)	0.05 ***	(0.01)
Any Service Learning (with control variables)		0.07 ***	(0.01)	0.07 ***	(0.01)
Service Learning 1		0.05 ***	(0.01)	0.05 ***	(0.01)
Service Learning 1 (with control variables)		0.05 ***	(0.01)	0.05 **	(0.02)
Service Learning 2		0.10 ***	(0.01)	0.08 ***	(0.01)
Service Learning 2 (with control variables)		0.07 ***	(0.01)	0.11 ***	(0.01)
Service Learning 3		0.09 ***	(0.03)	0.09 ***	(0.03)
Service Learning 3 (with control variables)		0.04	(0.03)	0.04	(0.03)
Multiple SL		0.02	(0.01)	0.05 ***	(0.01)
Multiple SL (with control variables)		0.02	(0.01)	0.04 *	(0.02)
Non-binary treatment (Reference group: 0 times)					
Any Service Learning:	Once	0.08 ***	(0.01)	0.06 ***	(0.01)
Any Service Learning:	Twice	0.05 ***	(0.01)	0.07 ***	(0.01)
Any Service Learning:	Three + times	0.14 ***	(0.02)	0.08 ***	(0.02)
Any Service Learning:	Once (with controls)	0.07 ***	(0.01)	0.06 ***	(0.01)
Any Service Learning:	Twice (with controls)	0.08 ***	(0.01)	0.09 *	(0.02)
Any Service Learning:	Three + times (with controls)	0.13 ***	(0.02)	0.09 **	(0.02)
Service Learning 1:	Once	0.07 ***	(0.02)	0.05 ***	(0.02)
Service Learning 1:	Twice	-0.01	(0.03)	-0.02	(0.03)
Service Learning 1:	Three + times	0.02	(0.05)	0.01	(0.05)
Service Learning 1:	Once (with controls)	0.06 ***	(0.02)	0.05 ***	(0.02)
Service Learning 1:	Twice (with controls)	0.03	(0.03)	0.04	(0.03)
Service Learning 1:	Three + times (with controls)	0.06	(0.05)	0.07	(0.05)
Service Learning 2:	Once	0.06 ***	(0.01)	0.06 ***	(0.01)
Service Learning 2:	Two + times	0.30 ***	(0.03)	0.26 ***	(0.03)
Service Learning 2:	Once (with controls)	0.04 **	(0.01)	0.09 ***	(0.02)
Service Learning 2:	Two + times (with controls)	0.22 ***	(0.03)	0.24 **	(0.03)

* p < .05, ** p < .01, *** p < .001.   Robust standard errors in parentheses.
Control variables: age in the first term, gender, race/ethnicity groups, Pell-ever status, ACT composite score, college in first term.
# Average Treatment Effect estimated with normalized Inverse Probability of Treatment Weights truncated at 1% and 99%.

Figure 8. Predicted increase in probability of transfer: Non-binary treatments



Note: If the 95% confidence interval crosses the vertical red line, the result is not statistically significant.

The third set of logistic models estimates the probability of student departure as a result of service learning participation. **Table 9** and **Figures 9** and **10** demonstrate the findings for this outcome for binary and non-binary treatments. We find that service learning participation has a strong negative statistically significant effect on the probability of student departure. The probability of departure is 20 percentage points lower for any service learning students than for their counterparts. The likelihood of departure is 19 percentage points lower for participants in duration level 1, 17 percentage points lower in duration level 2, and 37 percentage points lower for completers of level 3 than for nonparticipants. Taking different service learning levels decrease the likelihood of departure by 18 percentage points. The dosage (frequency) analysis shows that the effect size grows with an increase in participation frequency. The probability of departure drops by 19 percentage points for students completing service learning once, 27 percentage points for students taking it twice, and 33 percentage points for students experiencing this HIP three or more times as compared to nonparticipants. Similar effects are observed for duration levels 1 and 2.

Table 9. Predicted decrease in probability of departure as a result of service learning participation

Service learning types, duration, and frequency	Unweighted Sample		Weighted Sample #	
	Margin	SE	Margin	SE
Binary treatment (Reference group: nonparticipants)				
Any Service Learning	- 0.14 ***	(0.01)	- 0.13 ***	(0.01)
Any Service Learning (with control variables)	- 0.17 ***	(0.01)	- 0.20 ***	(0.01)
Service Learning 1	- 0.12 ***	(0.01)	- 0.14 ***	(0.01)
Service Learning 1 (with control variables)	- 0.16 ***	(0.02)	- 0.19 ***	(0.02)
Service Learning 2	- 0.16 ***	(0.01)	- 0.12 ***	(0.01)
Service Learning 2 (with control variables)	- 0.12 ***	(0.01)	- 0.17 ***	(0.02)
Service Learning 3	- 0.45***	(0.05)	- 0.45 ***	(0.05)
Service Learning 3 (with control variables)	- 0.37 ***	(0.05)	- 0.37 ***	(0.04)
Multiple SL	- 0.04 **	(0.01)	- 0.12 ***	(0.02)
Multiple SL (with control variables)	- 0.05 *	(0.01)	- 0.18 ***	(0.02)
Non-binary treatment (Reference group: 0 times)				
Any Service Learning: Once	- 0.16 ***	(0.01)	- 0.13 ***	(0.01)
Any Service Learning: Twice	- 0.09 ***	(0.01)	- 0.14 ***	(0.01)
Any Service Learning: Three + times	- 0.27 ***	(0.02)	- 0.22 ***	(0.03)
Any Service Learning: Once (with controls)	- 0.17 ***	(0.01)	- 0.19 ***	(0.01)
Any Service Learning: Twice (with controls)	- 0.16 ***	(0.01)	- 0.27 ***	(0.02)
Any Service Learning: Three + times (with controls)	- 0.30 ***	(0.02)	- 0.33 ***	(0.02)
Service Learning 1: Once	- 0.14 ***	(0.02)	- 0.12 ***	(0.02)
Service Learning 1: Twice	- 0.07 *	(0.03)	- 0.06 *	(0.03)
Service Learning 1: Three + times	- 0.18 ***	(0.05)	- 0.20 ***	(0.05)
Service Learning 1: Once (with controls)	- 0.15 ***	(0.02)	- 0.18 ***	(0.02)
Service Learning 1: Twice (with controls)	- 0.20 ***	(0.03)	- 0.26 ***	(0.03)
Service Learning 1: Three + times (with controls)	- 0.31 ***	(0.04)	- 0.38 ***	(0.04)
Service Learning 2: Once	- 0.12 ***	(0.01)	- 0.09 ***	(0.01)
Service Learning 2: Two + times	- 0.35 ***	(0.02)	- 0.33 ***	(0.03)
Service Learning 2: Once (with controls)	- 0.09 ***	(0.01)	- 0.15 ***	(0.02)
Service Learning 2: Two + times (with controls)	- 0.28 ***	(0.03)	- 0.34 ***	(0.03)

* p < .05, ** p < .01, *** p < .001. Robust standard errors in parentheses.
Control variables: age in the first term, gender, race/ethnicity groups, Pell-ever status, ACT composite score, college in first term.
# Average Treatment Effect estimated with normalized Inverse Probability of Treatment Weights truncated at 1% and 99%.

Figure 9. Predicted decrease in probability of departure: Binary treatments

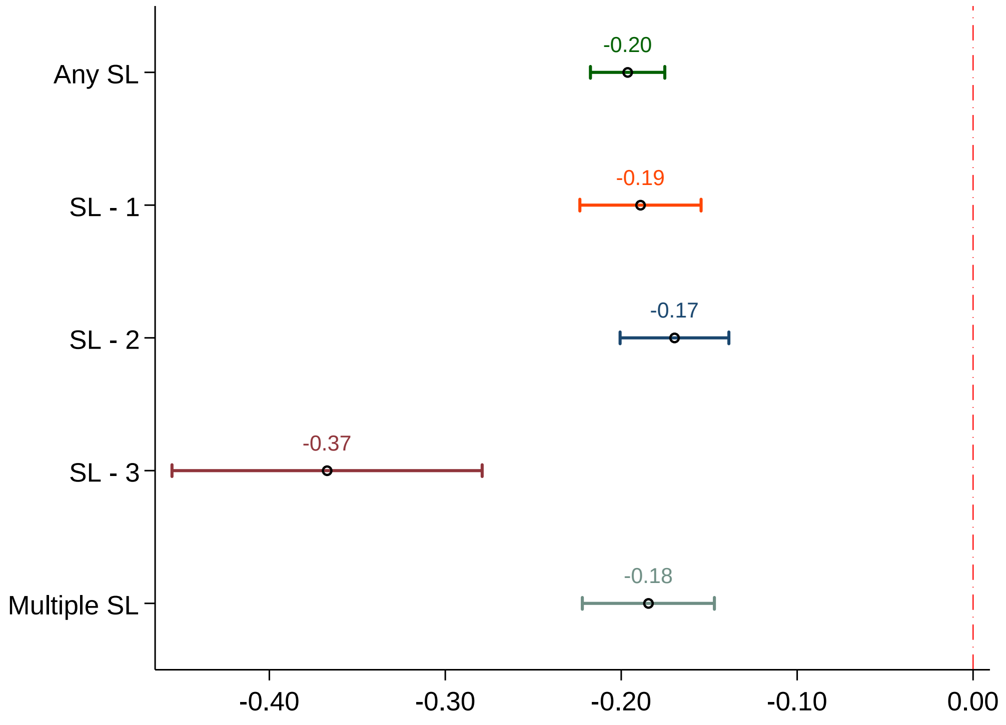
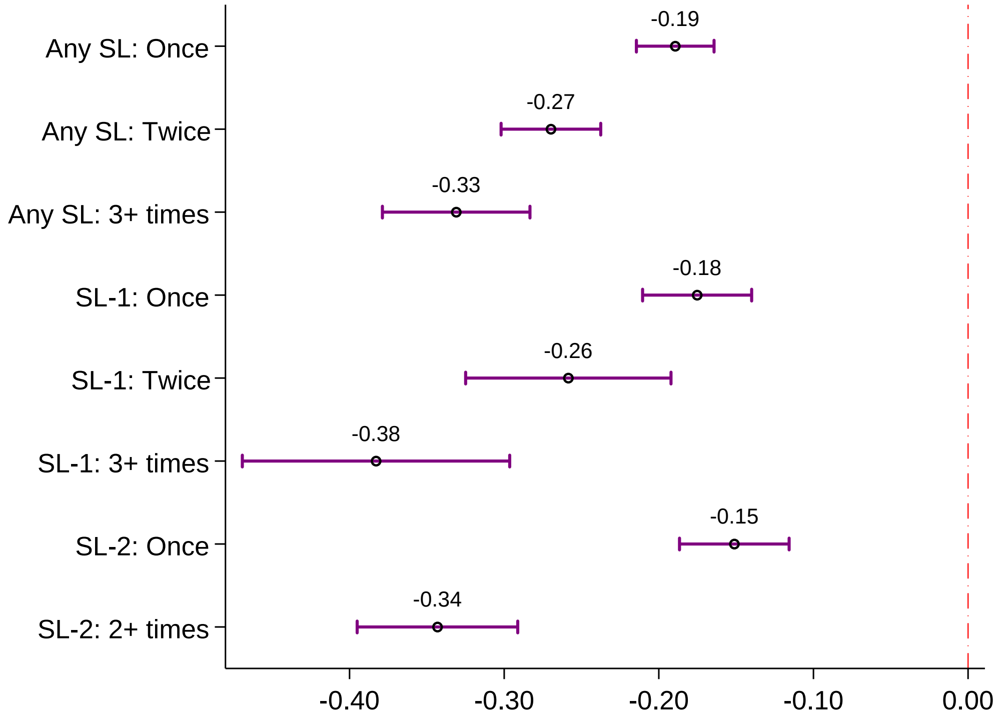


Figure 10. Predicted decrease in probability of departure: Non-binary treatments



The next set of models estimates the effect of service learning participation on final cumulative GPA (**Table 10** and **Figures 11** and **12**). In every model, all coefficients are highly statistically significant and positive. For binary treatments in the doubly robust models, the estimated impact of service learning participation on GPA ranges from 0.26 (*Multiple Service Learning*) to 0.60 (*Service Learning 3*). For multiple-level treatments (Figure 12), the effect ranges from 0.24 (*Service Learning 2*, completed once) to 0.66 (*Service Learning 1*, taken three or more times). We find that the estimated effect on GPA grows with an increase in frequency of service learning participation. Completing service learning once is expected to increase final GPA by 0.31 points, on average; taking it twice leads to an increase of 0.39 points; and participating in service learning experiences three or more times is associated with about half a point increase in the final GPA as compared to nonparticipants. Similar increases are observed by duration level. All findings must be interpreted in the context of sample sizes of the respective analytic samples.

Figure 11. Estimated effect on final GPA: Binary treatments

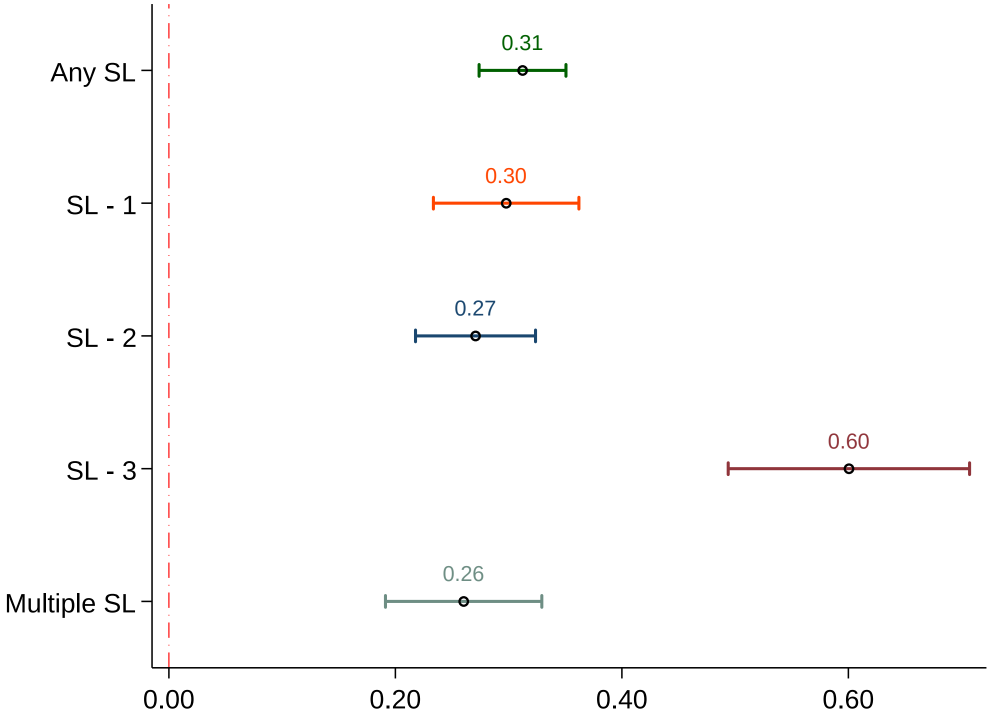
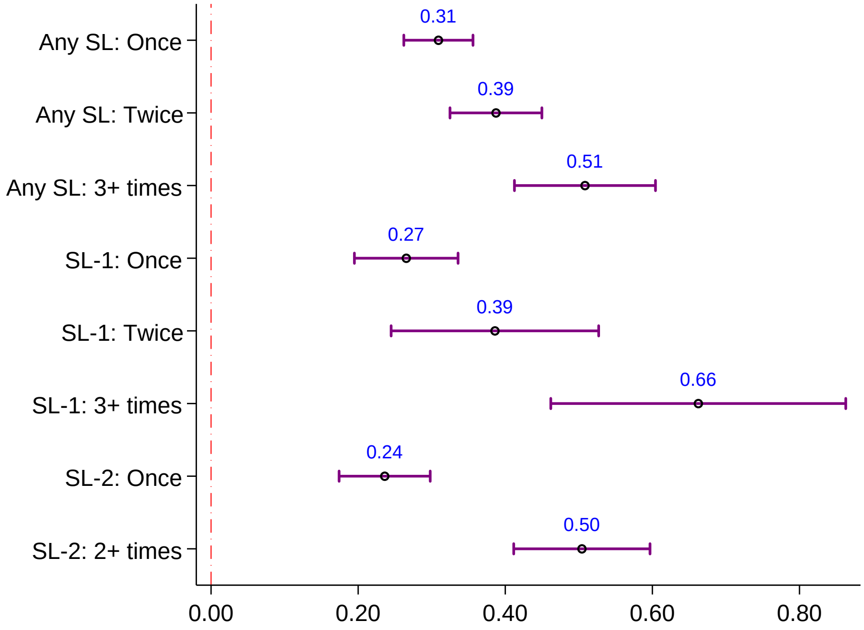


Table 10. OLS estimates of impact of service learning participation on final GPA

Service learning types, duration, and frequency		Unweighted Sample		Weighted Sample #	
		β	SE	β	SE
Binary treatment (Reference group: nonparticipants)					
Any Service Learning		0.29 ***	(0.02)	0.26 ***	(0.02)
Any Service Learning (with control variables)		0.26 ***	(0.02)	0.31 ***	(0.02)
Service Learning 1		0.21 ***	(0.03)	0.22 ***	(0.03)
Service Learning 1 (with control variables)		0.22 ***	(0.03)	0.30 ***	(0.03)
Service Learning 2		0.34 ***	(0.02)	0.28 ***	(0.02)
Service Learning 2 (with control variables)		0.20 ***	(0.02)	0.27 ***	(0.03)
Service Learning 3		0.78 ***	(0.06)	0.78 ***	(0.06)
Service Learning 3 (with control variables)		0.60 ***	(0.05)	0.60 ***	(0.05)
Multiple SL		0.09 ***	(0.03)	0.24 ***	(0.03)
Multiple SL (with control variables)		0.07 *	(0.03)	0.26 ***	(0.04)
Non-binary treatment (Reference group: 0 times)					
Any Service Learning:	Once	0.33 ***	(0.02)	0.24 ***	(0.02)
Any Service Learning:	Twice	0.18 ***	(0.02)	0.25 ***	(0.03)
Any Service Learning:	Three + times	0.52 ***	(0.05)	0.40 ***	(0.05)
Any Service Learning:	Once (with controls)	0.26 ***	(0.02)	0.31 ***	(0.02)
Any Service Learning:	Twice (with controls)	0.23 ***	(0.03)	0.39 ***	(0.03)
Any Service Learning:	Three + times (with controls)	0.46 ***	(0.04)	0.51 ***	(0.05)
Service Learning 1:	Once	0.23 ***	(0.03)	0.17 ***	(0.03)
Service Learning 1:	Twice	0.10	(0.06)	0.08	(0.06)
Service Learning 1:	Three + times	0.31 ***	(0.10)	0.35 **	(0.10)
Service Learning 1:	Once (with controls)	0.21 ***	(0.03)	0.27 ***	(0.04)
Service Learning 1:	Twice (with controls)	0.24 ***	(0.07)	0.39 ***	(0.07)
Service Learning 1:	Three + times (with controls)	0.44 ***	(0.10)	0.66 ***	(0.10)
Service Learning 2:	Once	0.26 ***	(0.03)	0.21 ***	(0.03)
Service Learning 2:	Two + times	0.65 ***	(0.04)	0.57 ***	(0.05)
Service Learning 2:	Once (with controls)	0.14 ***	(0.03)	0.24 ***	(0.03)
Service Learning 2:	Two + times (with controls)	0.44 ***	(0.04)	0.50 ***	(0.05)

* p < .05, ** p < .01, *** p < .001. Robust standard errors in parentheses.
Control variables: age in the first term, gender, race/ethnicity groups, Pell-ever status, ACT composite score, college in first term.
# Average Treatment Effect estimated with normalized Inverse Probability of Treatment Weights truncated at 1% and 99%.

Figure 12. Estimated effect on final GPA: Non-binary treatments



The EHA models estimate hazards for three outcomes of interest: graduation, transfer to a four-year institution, and student departure. Unlike the probability of outcome in the logistic models, hazards for graduation, transfer, and departure are estimated based on both whether the respective event took place and how long it took students to experience it. Therefore, interpretation of outcomes in the logistic and EHA models is different. Similar to the logistic models, the EHA results are presented both for binary and non-binary treatment variables in the same table although different methods were used in each case.

Table 11 and **Figures 13** and **14** provide estimates of the hazard for graduation. We find that students participating in any service learning face a 14-percent higher hazard for graduation (they are 14 percent more likely to graduate in any semester) than nonparticipants. The hazard is 19 percent higher for completers of *Service Learning 2* and 62 percent higher for completers of *Service Learning 3* than for non-completers. For non-binary treatments, participation in *Any Service Learning* once increases the completion hazard by 15 percent and twice by 31 percent. Completing *Service Learning 2* once and two or more times raises the graduation hazard by 18 and 48 percent, respectively. These findings indicate that service learning participants are more likely to earn an award than nonparticipants in any semester, and graduation happens faster with longer duration and higher frequency of service learning completion.

Table 11. Cox proportional hazards model for time to graduation by service learning participation

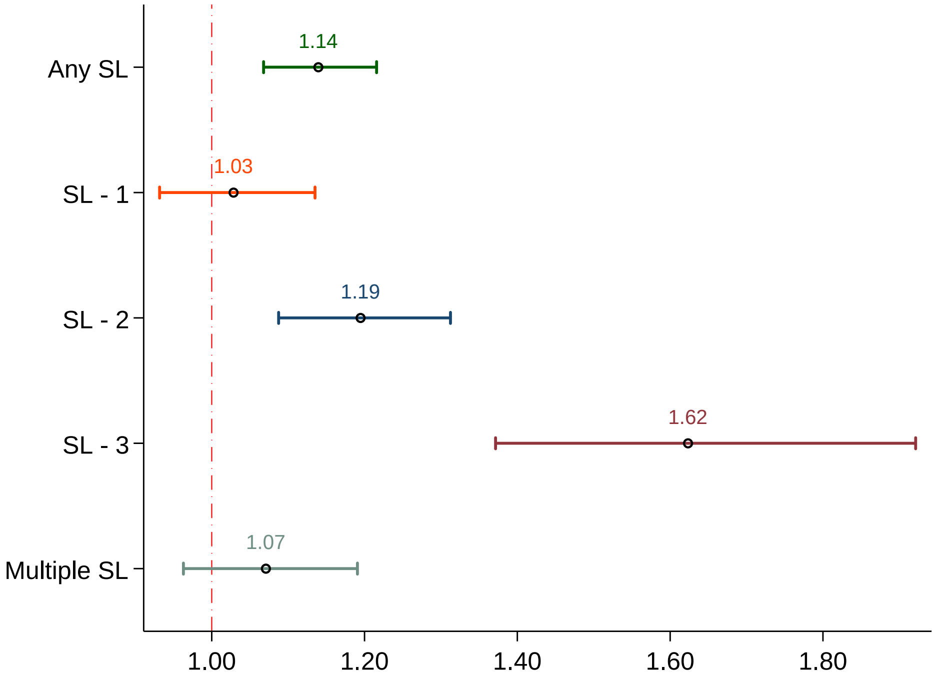
Service learning types, duration, and frequency		Unweighted Sample		Weighted Sample #	
		Hazard ratio	95% CI	Hazard ratio	95% CI
Binary treatment (Reference group: nonparticipants)					
Any Service Learning		1.25 *	[1.19, 1.31]	1.13 *	[1.07, 1.19]
Any Service Learning (with control variables)		1.13 *	[1.06, 1.20]	1.14 *	[1.07, 1.22]
Service Learning 1		1.07	[0.99, 1.16]	0.97	[0.89, 1.06]
Service Learning 1 (with control variables)		1.07	[0.98, 1.17]	1.03	[0.93, 1.14]
Service Learning 2		1.25 *	[1.17, 1.34]	1.14 *	[1.06, 1.22]
Service Learning 2 (with control variables)		1.12 *	[1.04, 1.21]	1.19 *	[1.09, 1.31]
Service Learning 3		1.91 *	[1.62, 2.25]	1.90 *	[1.62, 2.24]
Service Learning 3 (with control variables)		1.62 *	[1.37, 1.92]	1.62 *	[1.37, 1.92]
Multiple SL		1.12 *	[1.04, 1.20]	1.19 *	[1.09, 1.30]
Multiple SL (with control variables)		0.94	[0.86, 1.02]	1.07	[0.96, 1.19]
Non-binary treatment (Reference group: 0 times)					
Any Service Learning:	Once	1.24 *	[1.16, 1.31]	1.06	[0.99, 1.13]
Any Service Learning:	Twice	1.26 *	[1.18, 1.35]	1.28 *	[1.18, 1.39]
Any Service Learning:	Three + times	1.28 *	[1.15, 1.44]	1.12	[0.98, 1.28]
Any Service Learning:	Once (with controls)	1.14 *	[1.07, 1.22]	1.15 *	[1.06, 1.24]
Any Service Learning:	Twice (with controls)	1.10 *	[1.01, 1.20]	1.31 *	[1.18, 1.45]
Any Service Learning:	Three + times (with controls)	1.10	[0.97, 1.25]	1.13	[0.97, 1.31]
Service Learning 1:	Once	1.06	[0.97, 1.15]	0.99	[0.90, 1.09]
Service Learning 1:	Twice	1.20	[1.00, 1.45]	1.19	[0.98, 1.45]
Service Learning 1:	Three + times	0.94	[0.71, 1.24]	1.04	[0.79, 1.36]
Service Learning 1:	Once (with controls)	1.06	[0.97, 1.17]	1.05	[0.94, 1.16]
Service Learning 1:	Twice (with controls)	1.19	[0.96, 1.48]	1.25	[0.99, 1.58]
Service Learning 1:	Three + times (with controls)	0.91	[0.66, 1.25]	1.06	[0.78, 1.44]
Service Learning 2:	Once	1.18 *	[1.09, 1.27]	1.04	[0.96, 1.14]
Service Learning 2:	Two + times	1.53 *	[1.35, 1.73]	1.42 *	[1.25, 1.60]
Service Learning 2:	Once (with controls)	1.07 *	[0.98, 1.17]	1.18 *	[1.06, 1.32]
Service Learning 2:	Two + times (with controls)	1.29 *	[1.13, 1.46]	1.48 *	[1.28, 2.71]

* Indicates statistically significant result (95% confidence interval does not include 1). Robust confidence intervals in brackets.

Control variables: age, gender, race/ethnicity groups, Pell status, ACT score, college of enrollment. Time-varying covariates in binary treatment models: age, gender, Black, Pell status, ACT score. Time-varying covariates in non-binary treatment models: age, gender, Black, Pell status, ACT score (based on the tests of proportional hazard assumption violation).

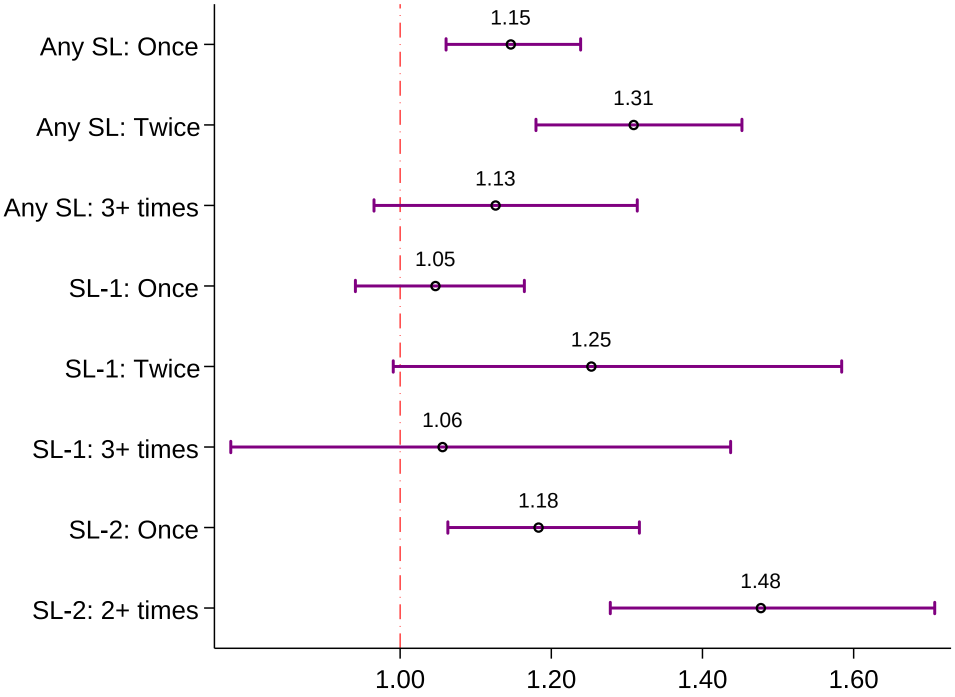
# Average Treatment Effect estimated with normalized Inverse Probability of Treatment Weights truncated at 1% and 99%.

Figure 13. Estimates of the effect on time to graduation: Binary treatments



Note: If the 95% confidence interval crosses the vertical red line, the result is not statistically significant.

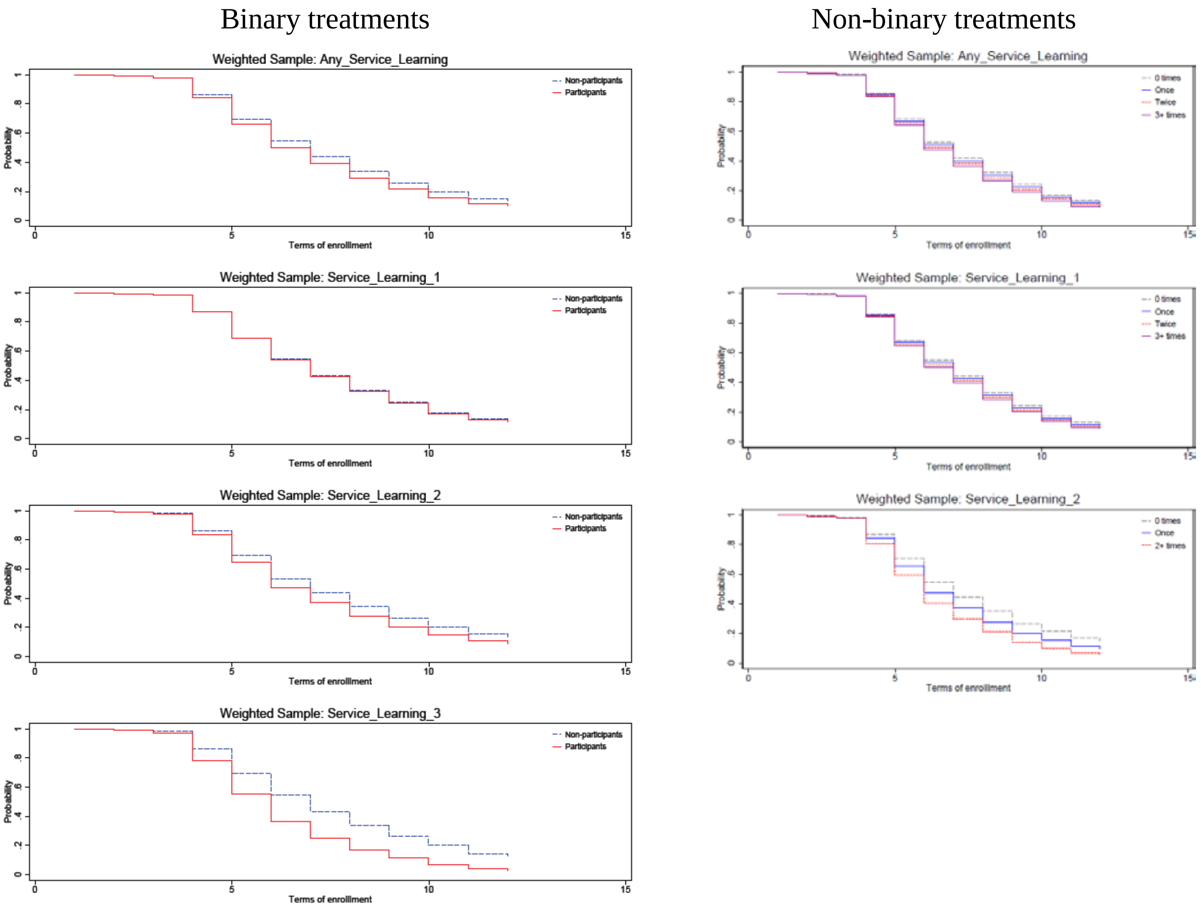
Figure 14. Estimates of the effect on time to graduation: Non-binary treatments



Note: If the 95% confidence interval crosses the vertical red line, the result is not statistically significant.

Figure 15 plots the survival curves for participants and nonparticipants using estimates from the weighted samples. The left side of the panel depicts survival curves for binary treatment variables (any service learning and three duration levels); the right side shows the change in survival probability for non-binary treatment variables. The survival functions are estimated at the mean values of other predictors in the model. In line with the above results, the curves indicate that service learning participants have a higher "risk" of graduation than nonparticipants: the survival curve for the former decreases faster than for the latter. The gap between the lines is smallest for *Service Learning 1*, for which the respective model did not find statistically significant results. The survivor functions for multiple levels of treatment for each duration level indicate that students with higher frequency of service learning participation are more likely to have graduated in a given semester than students with lower frequency and nonparticipants.

Figure 15. Estimated survival curves by service learning participation: Graduation, weighted samples

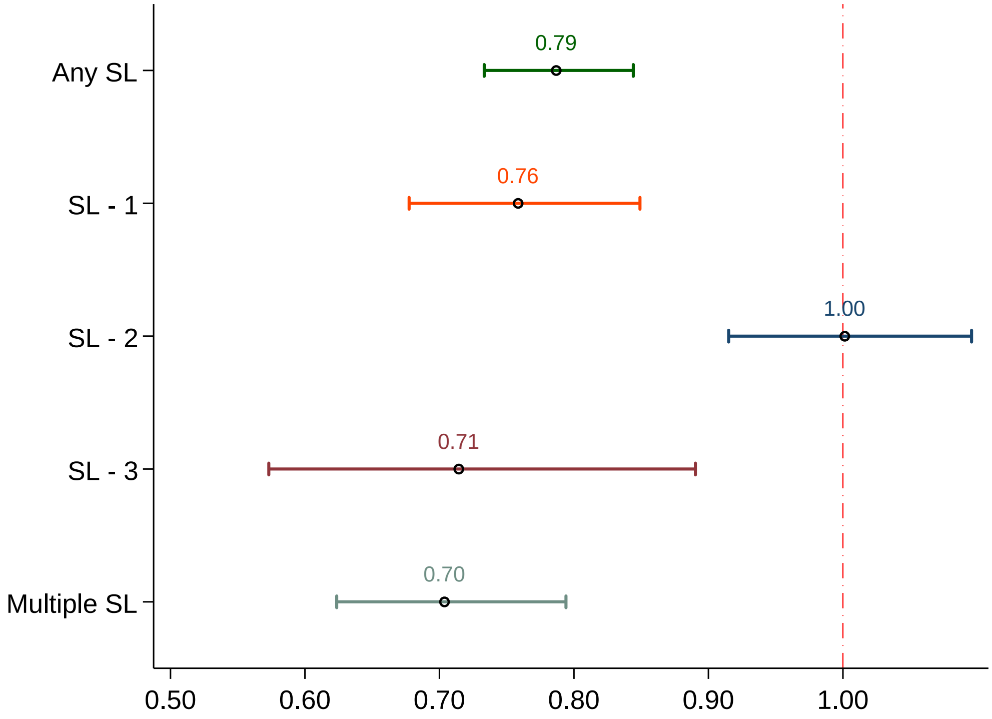

Binary treatments

Non-binary treatments

Note: Graphs depict estimated survivor functions for the time-to-graduation models at the mean value of other predictors.

The next set of models estimates hazards for university transfer. **Table 12** includes all estimates from these specifications, and **Figures 16** and **17** depict the results graphically. Contrary to expectations, we find that participation in service learning decreases hazard for transfer to university (i.e., participants progress to transfer slower than nonparticipants). Namely, the decrease in transfer hazard is 21 percent for *Any Service Learning*, 24 percent for *Service Learning 1*, 29 percent for *Service Learning 3*, and 30 percent for *Multiple Service Learning*. Participation in the duration level *Service Learning 2* is not found to exert an identifiable effect on this outcome. In the dosage analysis models (Figure 17), we find that taking part in *Any Service Learning* once decreases hazard for transfer by 21 percent, twice by 27 percent, three or more times by 46 percent. A similar pattern is observed for *Service Learning 1*: completing it once, twice, and three or more times decreases the transfer hazard by 20, 36, and 56 percent, respectively. No statistically significant effects are observed for the duration level *Service Learning 2*.

Figure 16. Estimates of the effect on time to transfer: Binary treatments



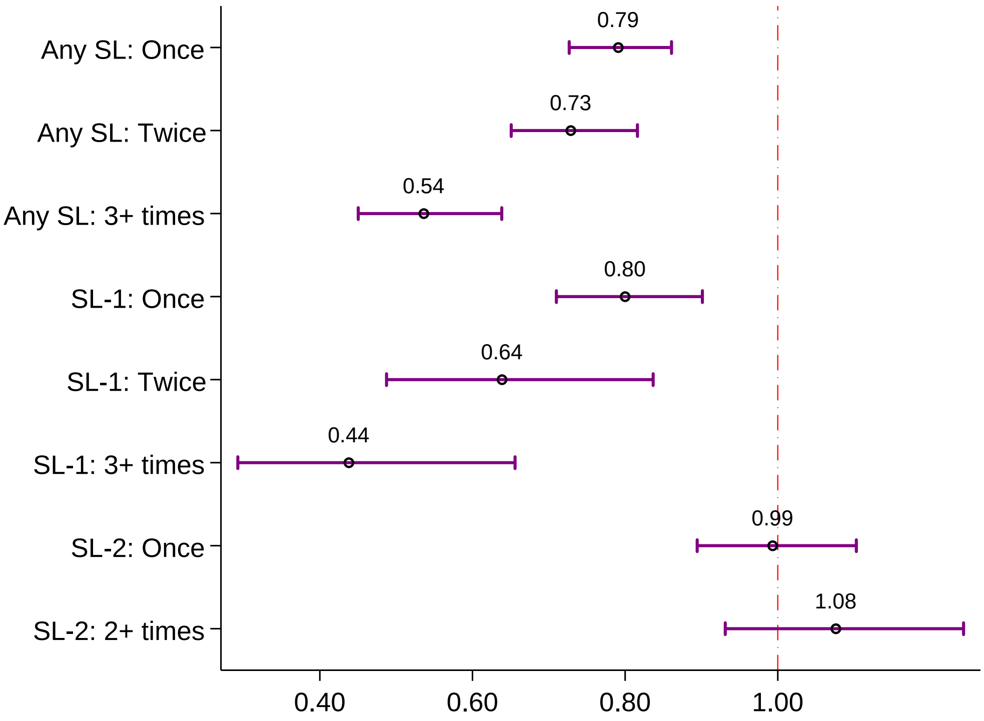
Note: If the 95% confidence interval crosses the vertical red line, the result is not statistically significant.

Table 12. Cox proportional hazards model for time to transfer by service learning participation

Service learning types, duration, and frequency		Unweighted Sample		Weighted Sample #	
		Hazard ratio	95% CI	Hazard ratio	95% CI
Binary treatment (Reference group: nonparticipants)					
Any Service Learning		0.95 *	[0.90, 1.00]	0.84 *	[0.79, 0.89]
Any Service Learning (with control variables)		0.86 *	[0.80, 0.91]	0.79 *	[0.73, 0.84]
Service Learning 1		0.92	[0.84, 1.01]	0.83 *	[0.75, 0.92]
Service Learning 1 (with control variables)		0.87 *	[0.79, 0.96]	0.76 *	[0.68, 0.85]
Service Learning 2		1.05	[0.97, 1.12]	0.98	[0.90, 1.06]
Service Learning 2 (with control variables)		1.01	[0.93, 1.10]	1.00	[0.92, 1.10]
Service Learning 3		0.80 *	[0.64, 1.00]	0.80 *	[0.64, 1.00]
Service Learning 3 (with control variables)		0.71 *	[0.57, 0.89]	0.71 *	[0.57, 0.89]
Multiple SL		0.91 *	[0.84, 0.99]	0.83 *	[0.75, 0.93]
Multiple SL (with control variables)		0.86 *	[0.78, 0.94]	0.70 *	[0.62, 0.79]
Non-binary treatment (Reference group: 0 times)					
Any Service Learning:	Once	0.96	[0.90, 1.03]	0.84 *	[0.78, 0.91]
Any Service Learning:	Twice	0.98	[0.91, 1.06]	0.90 *	[0.82, 0.99]
Any Service Learning:	Three + times	0.80 *	[0.70, 0.92]	0.67 *	[0.57, 0.78]
Any Service Learning:	Once (with controls)	0.87 *	[0.81, 0.94]	0.79 *	[0.73, 0.86]
Any Service Learning:	Twice (with controls)	0.89 *	[0.81, 0.97]	0.73 *	[0.65, 0.82]
Any Service Learning:	Three + times (with controls)	0.65 *	[0.56, 0.76]	0.54 *	[0.45, 0.64]
Service Learning 1:	Once	0.95	[0.86, 1.06]	0.88 *	[0.79, 0.98]
Service Learning 1:	Twice	0.92	[0.73, 1.16]	0.90	[0.71, 1.16]
Service Learning 1:	Three + times	0.65 *	[0.45, 0.93]	0.63 *	[0.43, 0.94]
Service Learning 1:	Once (with controls)	0.90	[0.81, 1.00]	0.80 *	[0.71, 0.90]
Service Learning 1:	Twice (with controls)	0.83	[0.65, 1.06]	0.64 *	[0.49, 0.84]
Service Learning 1:	Three + times (with controls)	0.55 *	[0.36, 0.86]	0.44 *	[0.29, 0.66]
Service Learning 2:	Once	0.99	[0.91, 1.08]	0.95	[0.86, 1.05]
Service Learning 2:	Two + times	1.21 *	[1.07, 1.36]	1.08	[0.95, 1.23]
Service Learning 2:	Once (with controls)	0.97	[0.88, 1.07]	0.99	[0.89, 1.10]
Service Learning 2:	Two + times (with controls)	1.14	[1.01, 1.30]	1.08	[0.93, 1.24]

* Indicates statistically significant result (95% confidence interval does not include 1). Robust confidence intervals in brackets.
Control variables: age, gender, race/ethnicity groups, Pell-ever status, ACT composite score, college of enrollment.
Time-varying covariates in binary treatment models: age, Hispanic, White, Pell status. Time-varying covariates in non-binary treatment models: age, Asian, Hispanic, White, Pell-ever status (based on the tests of proportional hazard assumption violation).
# Average Treatment Effect estimated with normalized Inverse Probability of Treatment Weights truncated at 1% and 99%.

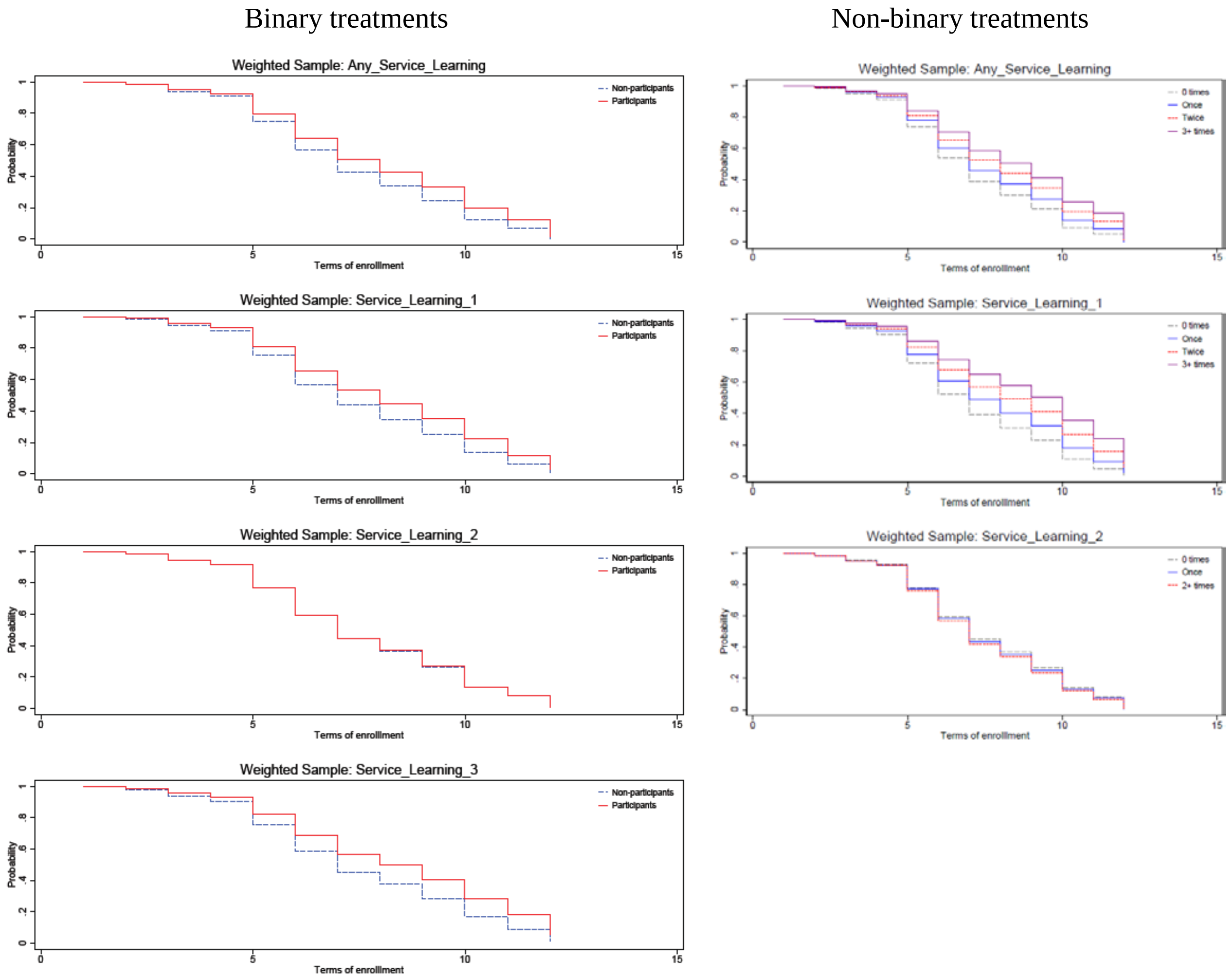
Figure 17. Estimates of the effect on time to transfer: Non-binary treatments



Note: If the 95% confidence interval crosses the vertical red line, the result is not statistically significant.

The above findings are corroborated by **Figure 18**, which plots survival curves for any service learning participants and completers of specific duration levels as well as survivor functions by frequency from the weighted samples. Similar to Figure 15, the left side of the panel shows survival curves for binary treatment variables, and the right side presents the change in survival probability for non-binary treatment variables. The survival functions are estimated at the mean values of other predictors in the model. Figure 18 shows that service learning students progress to university transfer at a slower rate than nonparticipants, and each additional frequency level slows down this progression. The gap between the lines is smallest for *Service Learning 2*, for which the respective models did not find statistically significant results. The *Discussion* section explains how this finding may be reconciled with the finding of the higher probability of transfer for service learning participants in the logistic models.

Figure 18. Estimated survival curves by service learning participation: Transfer, weighted samples



Note: Graphs depict estimated survivor functions for the time-to-transfer models at the mean value of other predictors.

The final set of models estimates the hazard for student departure. **Table 13** and **Figures 19** and **20** provide hazard ratios for the respective specifications. We find that service learning participation has a strong negative statistically significant effect on progression to departure. For binary treatment variables, the hazard for departure decreases for participants as compared to nonparticipants by 47 percent for *Any Service Learning* , 48 percent for *Service Learning 1*, 41 percent for *Service Learning 2*, 73 percent for *Service Learning 3*, and 45 percent for *Multiple Service Learning*. For nonbinary treatment, the effect size increases with the growth in frequency of experience and ranges from 37 percent (*Service Learning 2*, once) to 78 percent (*Service Learning 1*, three or more times) decline in the hazard for transfer.

Table 13. Cox proportional hazards model for time to departure by service learning participation

Service learning types, duration, and frequency		Unweighted Sample		Weighted Sample #	
		Hazard ratio	95% CI	Hazard ratio	95% CI
Binary treatment (Reference group: nonparticipants)					
Any Service Learning		0.67 *	[0.64, 0.69]	0.68 *	[0.65, 0.71]
Any Service Learning (with control variables)		0.57 *	[0.54, 0.60]	0.53 *	[0.50, 0.56]
Service Learning 1		0.69 *	[0.64, 0.75]	0.65 *	[0.60, 0.71]
Service Learning 1 (with control variables)		0.59 *	[0.54, 0.64]	0.52 *	[0.47, 0.58]
Service Learning 2		0.64 *	[0.60, 0.68]	0.70 *	[0.65, 0.75]
Service Learning 2 (with control variables)		0.68 *	[0.63, 0.73]	0.59 *	[0.54, 0.64]
Service Learning 3		0.24 *	[0.17, 0.33]	0.24 *	[0.17, 0.33]
Service Learning 3 (with control variables)		0.27 *	[0.20, 0.38]	0.27 *	[0.20, 0.38]
Multiple SL		0.89 *	[0.84, 0.95]	0.70 *	[0.64, 0.76]
Multiple SL (with control variables)		0.82 *	[0.76, 0.89]	0.55 *	[0.50, 0.61]
Non-binary treatment (Reference group: 0 times)					
Any Service Learning:	Once	0.64 *	[0.60, 0.68]	0.68 *	[0.64, 0.72]
Any Service Learning:	Twice	0.78 *	[0.73, 0.83]	0.66 *	[0.61, 0.71]
Any Service Learning:	Three + times	0.42 *	[0.37, 0.49]	0.47 *	[0.40, 0.54]
Any Service Learning:	Once (with controls)	0.58 *	[0.55, 0.62]	0.53 *	[0.50, 0.58]
Any Service Learning:	Twice (with controls)	0.60 *	[0.56, 0.65]	0.41 *	[0.37, 0.45]
Any Service Learning:	Three + times (with controls)	0.34 *	[0.30, 0.39]	0.29 *	[0.25, 0.34]
Service Learning 1:	Once	0.68 *	[0.62, 0.74]	0.70 *	[0.64, 0.77]
Service Learning 1:	Twice	0.83 *	[0.71, 0.96]	0.84 *	[0.72, 0.99]
Service Learning 1:	Three + times	0.55 *	[0.42, 0.73]	0.52 *	[0.39, 0.71]
Service Learning 1:	Once (with controls)	0.62 *	[0.56, 0.68]	0.54 *	[0.48, 0.60]
Service Learning 1:	Twice (with controls)	0.54 *	[0.45, 0.64]	0.39 *	[0.32, 0.49]
Service Learning 1:	Three + times (with controls)	0.34 *	[0.25, 0.45]	0.22 *	[0.16, 0.31]
Service Learning 2:	Once	0.73 *	[0.68, 0.78]	0.77 *	[0.72, 0.84]
Service Learning 2:	Two + times	0.31 *	[0.25, 0.37]	0.32 *	[0.26, 0.40]
Service Learning 2:	Once (with controls)	0.76 *	[0.70, 0.82]	0.63 *	[0.57, 0.69]
Service Learning 2:	Two + times (with controls)	0.35 *	[0.29, 0.43]	0.28 *	[0.23, 0.35]

* Indicates statistically significant result (95% confidence interval does not include 1). Robust confidence intervals in brackets. Control variables: age, gender, race/ethnicity groups, Pell status, ACT score, college of enrollment. Time-varying covariates in binary treatment models: age, gender, Asian, Black, Pell-ever status. Time-varying covariates in non-binary treatment models: age, gender, Asian, Black, Hispanic, ACT score (based on the tests of proportional hazard assumption violation).
# Average Treatment Effect estimated with normalized Inverse Probability of Treatment Weights truncated at 1% and 99%.

Figure 19. Estimates of the effect on time to departure: Binary treatments

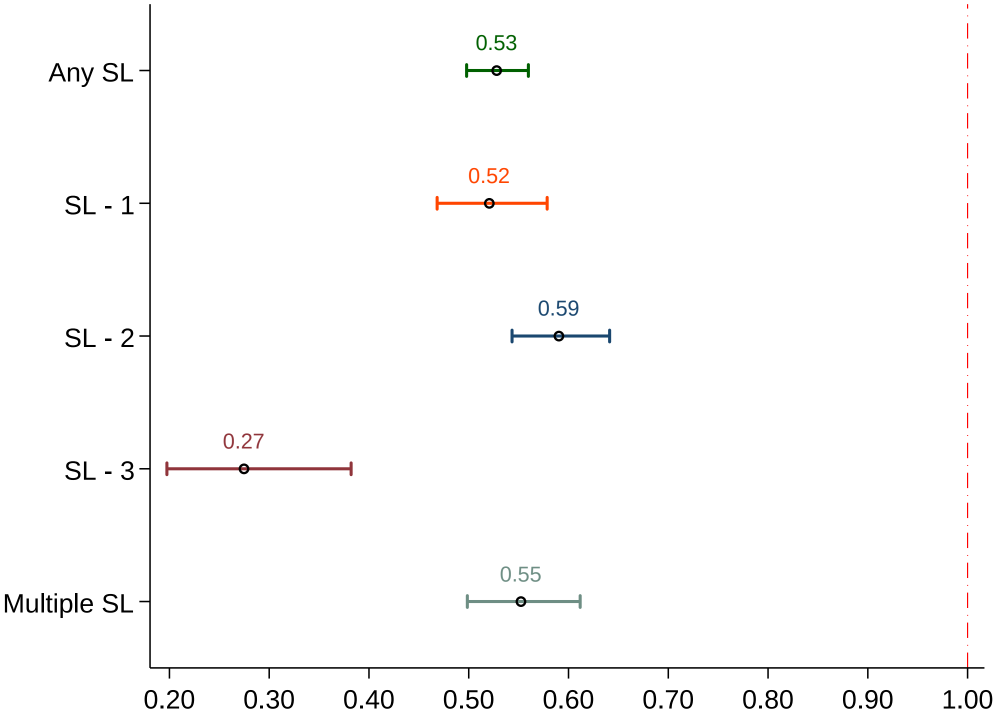


Figure 20. Estimates of the effect on time to departure: Non-binary treatments

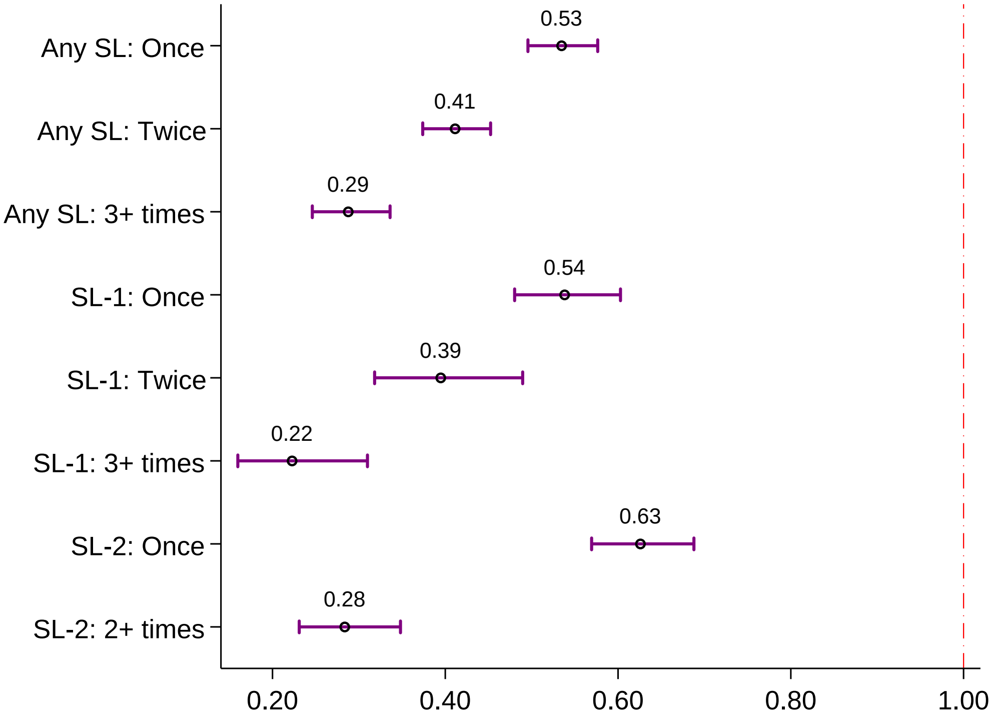


Table 14. The effect of service learning participation on all outcomes: Summary of statistically significant results, final models

	Increase in GPA (points)	Change in predicted probability of:			Increase in hazard for graduation	Decrease in hazard for:	
		Graduation	Transfer	Departure		Transfer	Departure
BINARY TREATMENT							
Any Service Learning	0.31	17 pp.	7 pp.	- 20 pp.	14%	21%	47%
Service Learning - 1	0.30	16 pp.	5 pp.	- 19 pp.	–	24%	48%
Service Learning - 2	0.27	15 pp.	11 pp.	- 17 pp.	19%	–	41%
Service Learning - 3	0.60	33 pp.	–	- 37 pp.	62%	29%	73%
Multiple Service Learning	0.26	17 pp.	4 pp.	- 18 pp.	–	30%	45%
NON-BINARY TREATMENT							
Any Service Learning: Once	0.31	16 pp.	6 pp.	- 19 pp.	15%	21%	47%
Any Service Learning: Twice	0.39	26 pp.	9 pp.	- 27 pp.	31%	27%	59%
Any Service Learning: 3+ times	0.51	31 pp.	9 pp.	- 33 pp.	–	46%	71%
Service Learning – 1: Once	0.27	13 pp.	5 pp.	- 18 pp.	–	20%	46%
Service Learning – 1: Twice	0.39	22 pp.	–	- 26 pp.	–	36%	61%
Service Learning – 1: 3+ times	0.66	34 pp.	–	- 38 pp.	–	56%	78%
Service Learning – 2: Once	0.24	14 pp.	9 pp.	- 15 pp.	18%	–	37%
Service Learning – 2: 2+ times	0.50	33 pp.	24 pp.	- 34 pp.	48%	–	72%

Note: Statistical significance is determined at p < .05. Refer to the respective tables in the *Results* section to obtain the specific significance level for each estimate.
Service Learning 3 was not used in frequency analysis (non-binary treatment) due to small sample size.
For *Service Learning 2*, the categories "*Twice*" and "*3+ times*" were combined into "*2+ times*" due to small size of the latter group.

Discussion

The study investigates the effect that service learning—a high impact practice (HIP) implemented at scale at Tennessee's Board of Regent's (TBR) community colleges—has on key college outcomes within twelve semesters of starting as first-time freshmen. Together with Iteration 1 (the first stage of the investigation), it represents the first attempt to quantitatively assess the unbiased effect of high impact practices at TBR institutions.[17] The prior descriptive and survey-based quantitative analyses—although comprehensive and large-scale—did not address the issue of student selection into HIP. We use several strategies to address the selectivity bias, identify the impact of service learning participation, and quantify its effects on completion, transfer, student departure, and academic performance. We find that students who participated in service learning have a higher probability of graduation and transfer to a four-year institution and a lower probability of departure, face a higher hazard for graduation and a lower hazard for transfer and departure, and tend to have a higher final GPA than their similar counterparts who did not participate in this high impact practice. The statistically significant effects also vary by service learning duration level and participation frequency during the observation period.

The study findings offer support to past evidence of positive academic effects of participating in high impact practices and contribute to the extant literature by using several strategies to examine the impact of a particular HIP on several college outcomes. First, we find an overwhelming support for the expectation that service learning participation increases the probability of earning a college credential and transferring to a four-year institution. For graduation—which is operationalized as earning either a technical certificate or a degree—statistically significant results are observed for all duration and frequency levels of service learning experiences. For university transfer, we find statistically significant impacts for any service learning participation as well as for particular duration and frequency levels. In community

[17] Iteration 1 of the study covered 9 semesters of tracking data (through summer 2020), while the current Iteration 2 analyzed 12 terms of data, through summer 2021.

college settings, both outcomes are measures of student and college success, and these findings attest to the efficacy of service learning as an impactful educational practice.

Second, participation in service learning is found to decrease time to graduation or, stated differently, to accelerate progression to completion. Statistically significant results are observed for participation in any service learning, but also for longer duration levels (10-19 hours and 20 or more hours of service), with a longer duration being associated with a higher effect size. In frequency analyses, the effects are found for any service learning and *Service Learning 2* duration (10-19 hours), with the effect growing with frequency. Based on these findings, we conclude that service learning is effective in propelling students towards graduation—especially if exposure to this experience is long enough and frequent enough.

Third, service learning is associated with a higher final GPA. Participants in any duration and frequency level tend to have significantly higher cumulative GPAs than their counterparts among similar nonparticipants. The estimates from the quantitative analysis are supported by the results of the descriptive analysis, which compared academic variables between the two groups. The employed methodology aimed to minimize the pre-existing differences among service learning participants and nonparticipants, including differences in academic abilities and preparation. Having tested the effectiveness of this bias-minimizing approach, we conclude that service learning has certain educational components that contribute to better grades in college. These findings provide an additional evidence that engagement in this HIP is an important predictor and driver of student success as measured by academic performance.

Fourth, students participating in service learning HIP are both less likely to depart (drop out or stop out) and demonstrate a longer time to departure than their similar counterparts among nonparticipants. Importantly, in this study student departure is interpreted as departing from higher education in general (based on all available data from the TBR data warehouse and the National Student Clearinghouse) and not just from the TBR community college system. The effect size increases with the growth in frequency of service learning experience. These findings indicate that service learning participation may contribute to creating conditions that provide for student retention. Although well documented in extant literature, this effect receives additional support in the current study, which accounts for the pre-existing differences between freshmen students who did and did not participate in service learning.

Finally, at first approximation, the findings for university transfer seem to contradict each other: in the logistic models, we find that service learning participation leads to a higher probability of transfer, while the Event History Analysis models produce lower hazards for transfer (indicating slower progression to this outcome). However, one needs to remember that these models focus on different outcomes: in logistic regression, the outcome is binary ("*Did a participant transfer?*"), whereas Event History Analysis models time to the event of transfer ("*Did a participant transfer? And if yes, how long did it take them to transfer?*"). Therefore, a possible interpretation of the quantitative analysis results is that service learning participation increases the probability of transfer but, simultaneously, delays progression to this outcome as compared to students who do not complete service learning.

The descriptive analysis of the sample's outcome (the *Dataset description* section) offers further evidence to this interpretation: it shows that a larger share of service learning participants transfer than nonparticipants (30.6% vs. 23%), but also that more participants do so after they earn a community college credential as compared to nonparticipants (20.3% vs. 11.4%). The average time to transfer is shorter for nonparticipants (5.26 semesters) than for any service learning participants (5.74 semesters) or participants in any duration level of this HIP. On average, service learning students also attempt more credit hours before they transfer than similar nonparticipants (65.0 vs. 57.9). All the above, and the results of the quantitative analysis and examination of the survivor function graphs, indicate that while service learning participants are more likely to transfer than their counterparts, they also take longer to transfer to a four-year institution. These findings are partly due to service learning participants attempting, on average, more credits (thus staying longer in community colleges than nonparticipants who transfer) and tending to earn a certificate or degree prior to transferring to university.

The study limitations are as follows: first, propensity score weighting and matching used thirty-three variables, which were selected based on theoretical considerations, prior research, knowledge about selection into HIPs, and data availability. Although tests show that these predictors and employed weighting and matching algorithms work well to address the selectivity bias, the bias is minimized only on the dimensions that are represented by these predictors, and the conditional independence assumption remains untestable (Angrist, 1997). Therefore, the conclusions depend on the following factors: predictors of service learning participation, assumptions underlying employed models, data availability and accuracy, and precision of the weighting and matching algorithms. In order to further control for the remaining differences among subjects, we employed doubly robust models with additional covariates.

Second, although sample size is large for most groups of service learning participants, it may be an issue for the following smaller categories: the service learning component of 20 or more hours of service (*Service Learning 3*) and frequency category "*Three or more times*." While we find statistically significant results for these categories in various models, these findings should be interpreted with caution because the true effect size is unknown due to larger confidence intervals than for other categories. (The results of the sensitivity analyses showed that the vast majority of findings were consistent across different models and approaches, although the estimated effect sizes varied.[18]) Next, the generalizability of the study is limited to first-time freshmen who participated in the service learning HIP as implemented at TBR community colleges as part of a large-scale reform effort in recent years. We do not make any claims about efficacy of service learning in other settings and contexts. Finally, the study examined the impact of service learning by duration and frequency; however, it did not use other, more refined, ways of classifying service learning experiences across TBR institutions. The potential interaction of service learning with other HIPs in affecting student and college success is also left for future investigations.

Some policy implications stem from this research. First, given the study's findings and limitations, we conclude that service learning at TBR is an effective intervention that contributes both to student educational success and community college's production function. As such, this high impact practice should be promoted in order to enhance opportunities of success for greater number of students. For example, making service learning a component of first-year experience—a practice already implemented

[18] As explained previously, sensitivity analyses were run in Iteration 1 of the study, which used 9 semesters of data.

in some colleges—seems a promising option of offering this opportunity to more students. Second, the results by duration level and frequency indicate that the respective impact may depend on how long and how often students participate in HIP. These findings may serve as an argument for implementing changes in the curriculum that, for example, may provide service learning opportunities of longer duration or higher frequency—depending on the desired outcome. To conclude, the study demonstrates the need to conduct more research on high impact practices to answer remaining questions about their effect.

Future investigations will examine the impact of other high impact practices implemented at TBR institutions and their potential interaction in promoting student middle-range and long-term success.[19] This research can be strengthened by examining the differences in service learning implementation across colleges, using more nuanced classifications of types of service learning experiences, and adding qualitative research components (focus groups, interviews, etc.) that will include analysis of students' perceptions and faculty inputs. One of the main questions that needs to be answered is the best timing for service learning experience (the one that leads to greater students success overall), and it should be addressed with both quantitative and qualitative research methods. As more data becomes available for analysis from the TBR institutions, the National Student Clearinghouse, the Tennessee Longitudinal Data System, and the Coleridge Initiative, it will become possible to examine additional college and labor market outcomes of HIP participants after their graduation and joining the workforce. The findings from this and future studies will contribute to promoting student success and meeting the goals of *Drive to 55* and other state and TBR initiatives.

[19] At the time of writing this report on Iteration 2 of the study, the TBR research team has completed a study on undergraduate research HIP (*The Impact of Undergraduate Research High Impact Practice on Community College Student Outcomes*), which is available at: https://www.tbr.edu/policy-strategy/presentations-and-papers

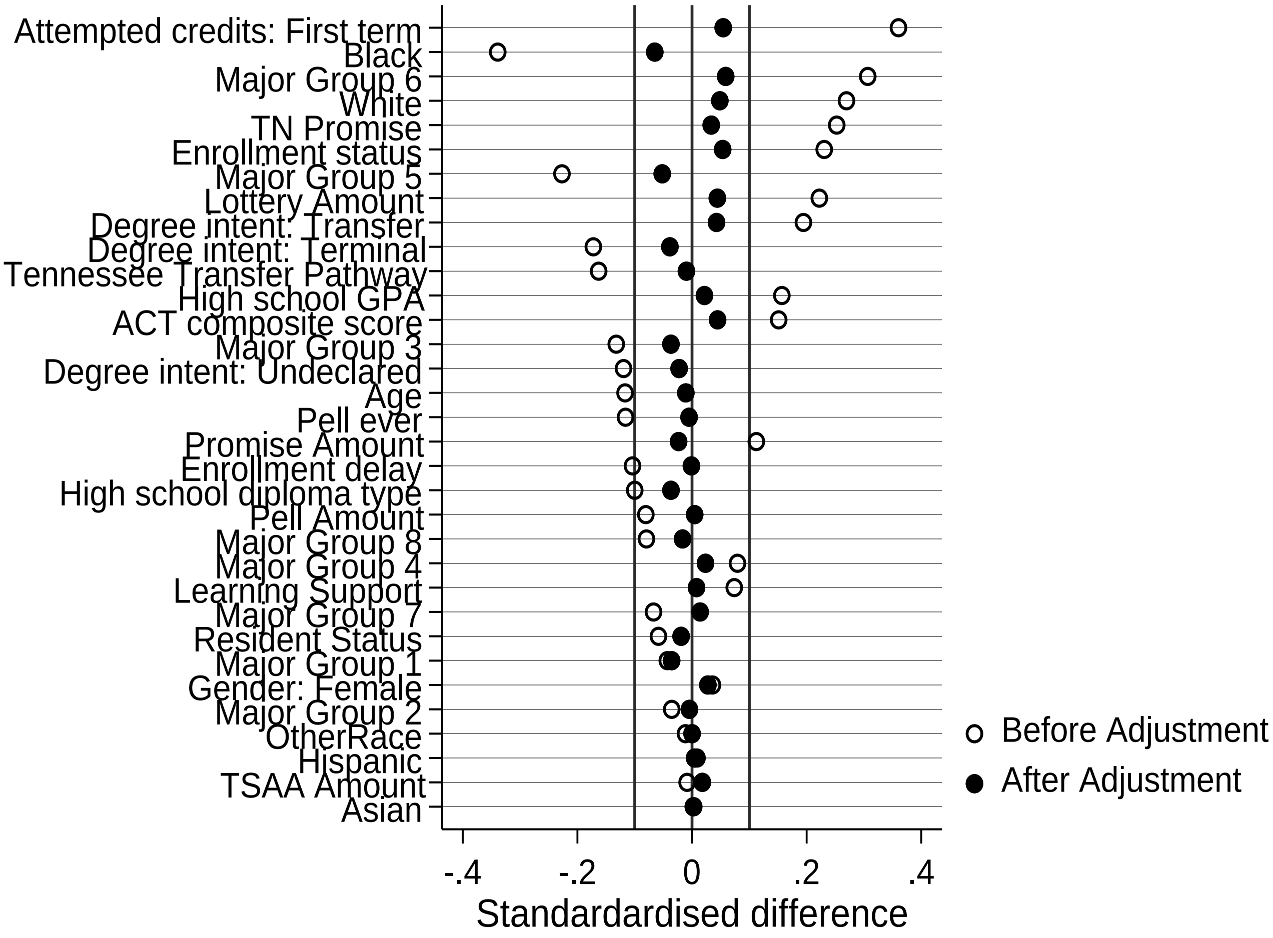
Appendices

Appendix 1. Standardized difference between mean values for treated and untreated groups before and after weighting, any service learning (binary treatment)

	Unweighted sample			Weighted sample		
	participants	non-participants	Stand. difference	participants	non-participants	Stand. difference
Age in the first semester	19.15	19.7	-0.117	19.53	19.59	-0.011
Gender: Female (0, 1)	0.57	0.56	0.036	0.58	0.56	0.028
Race/ethn.: Asian (0, 1)	0.01	0.01	0.003	0.01	0.01	0.002
Race/ethn.: Black (0, 1)	0.09	0.21	-0.339	0.16	0.18	-0.065
Race/ethn.: Hispanic (0, 1)	0.06	0.06	0.009	0.06	0.06	0.005
Race/ethn.: White (0, 1)	0.79	0.67	0.27	0.72	0.7	0.049
Race/ethn.: Other (0, 1))	0.04	0.05	-0.011	0.05	0.05	0
Learning Support (0, 1)	0.67	0.63	0.074	0.64	0.64	0.008
ACT composite score	19.31	18.73	0.151	19.03	18.86	0.045
High school GPA	3.12	2.99	0.157	3.04	3.02	0.022
Resident Status	1.02	1.03	-0.059	1.02	1.03	-0.019
High school diploma type	1.83	1.91	-0.1	1.87	1.9	-0.037
Enrollment delay	1.01	1.46	-0.104	1.36	1.37	-0.001
Pell Amount	1,296.56	1,403.31	-0.081	1,403.37	1,397.16	0.005
Promise Amount	411.89	338.39	0.112	329.91	345.22	-0.023
Lottery Amount	800.87	627.82	0.222	699.74	665.27	0.044
TSAA Amount	219.13	221.7	-0.008	229.2	223.73	0.018
Pell-eligible ever (0, 1)	0.6	0.66	-0.116	0.65	0.65	-0.005
Tennessee Promise (0, 1)	0.72	0.6	0.253	0.64	0.63	0.034
Degree intent: Transfer	0.71	0.62	0.194	0.66	0.64	0.043
Degree intent: Terminal	0.28	0.36	-0.172	0.33	0.35	-0.039
Degree intent: Undeclared	0	0.01	-0.119	0.01	0.01	-0.022
Enrollment status: Term 1	0.95	0.88	0.231	0.91	0.9	0.053
Attempted credits: Term 1	13.91	13.02	0.36	13.34	13.2	0.054
Transfer Pathway (0, 1)	0.15	0.22	-0.163	0.2	0.2	-0.01
Major Group 1 (0, 1)	0	0.01	-0.043	0	0.01	-0.036
Major Group 2 (0, 1)	0.03	0.03	-0.035	0.03	0.03	-0.004
Major Group 3 (0, 1)	0.08	0.12	-0.132	0.1	0.11	-0.037
Major Group 4 (0, 1)	0.04	0.03	0.079	0.03	0.03	0.024
Major Group 5 (0, 1)	0.1	0.17	-0.227	0.14	0.16	-0.052
Major Group 6 (0, 1)	0.61	0.46	0.307	0.52	0.49	0.059
Major Group 7 (0, 1)	0.05	0.07	-0.067	0.07	0.07	0.014
Major Group 8 (0, 1)	0.09	0.11	-0.08	0.1	0.1	-0.016

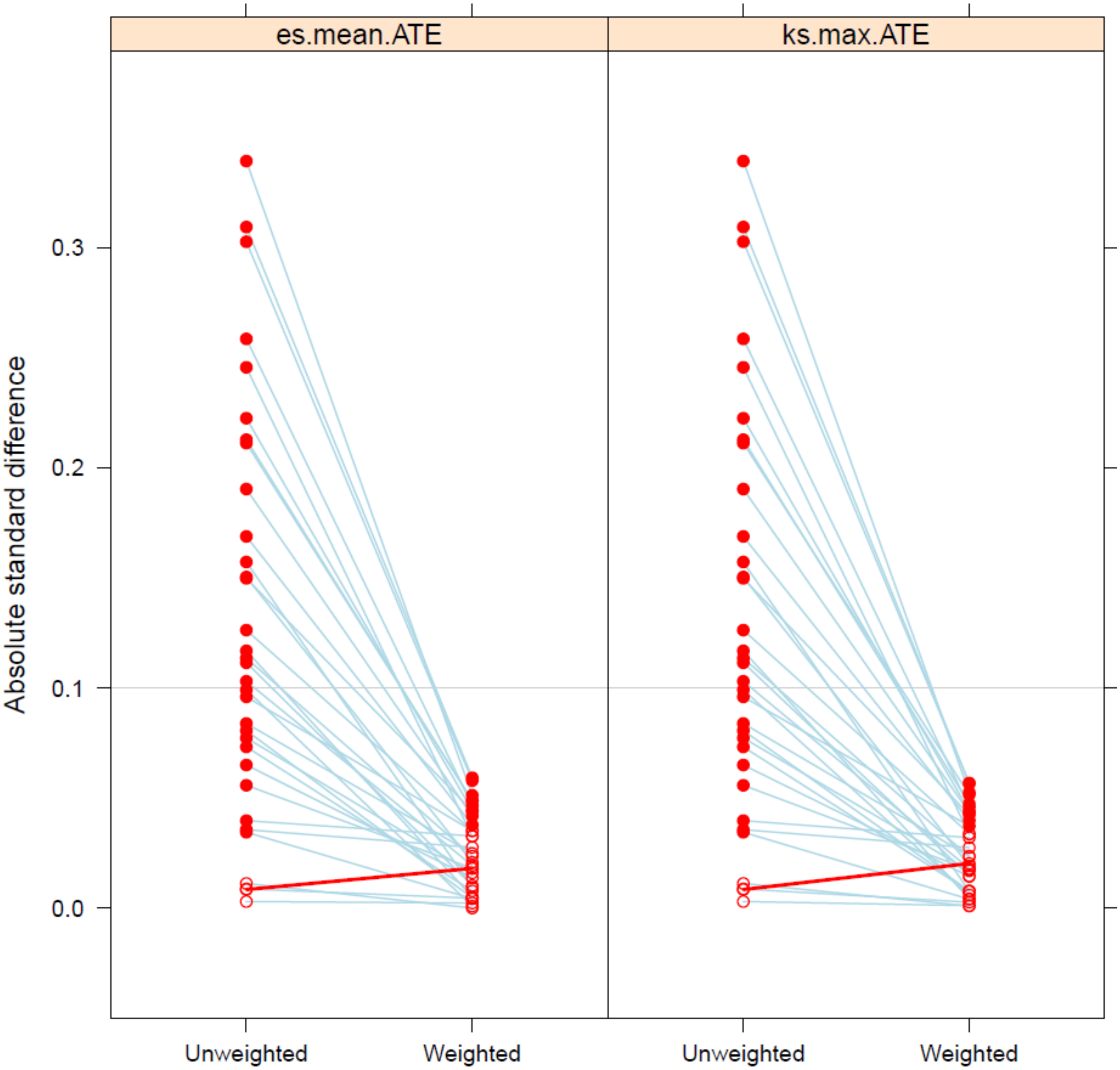
Note: An absolute standardized difference of ≤ 0.10 is considered balanced.

Appendix 2. Standardized difference between treated and untreated groups after ATE weighting, any service learning (binary treatment)



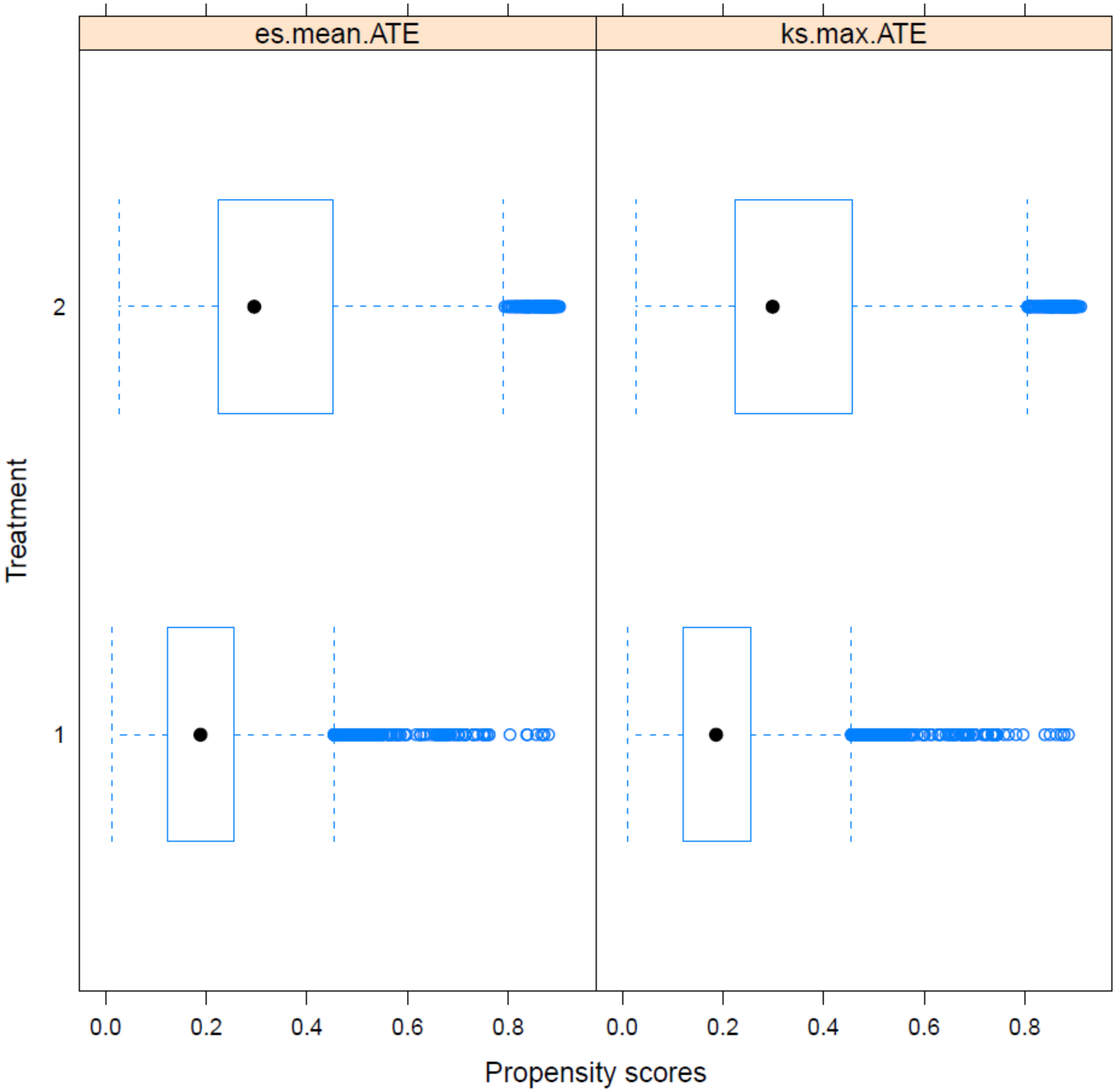
Note: An absolute standardized difference of ≤ 0.10 is considered balanced.

Appendix 3. The reduction in standardized differences between treated and untreated groups due to ATE weighting, any service learning (binary treatment)



Note: A (absolute) standardized difference of ≤ 0.10 is considered balanced.

Appendix 4. Boxplot of propensity scores



Appendix 5. Predicted increase in the probability of graduation: Model comparison, weighted and matched samples, 9-term tracking (fall 2007–summer 2020)

Estimand / PSA type	ATE / Inverse Probability of Treatment W.			ATT	ATT / Propensity Score Matching			
PSA model	GBR Modeling	Boosted regres.	Logistic	Kernel	Neighbors: 6	Neighbors: 4	Neighbors: 2	1-to-1 match
Any Service Learning	0.11	0.18	0.07	0.07	0.07	0.07	0.07	0.07
SE	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Any SL (controls)	0.15	0.17	0.12	0.12	0.12	0.11	0.11	0.11
SE	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Service Learning 1	0.09	0.08	0.05	0.05	0.05	0.05	0.05	0.07
SE	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02
p-value	0.00	0.00	0.00	0.01	0.01	0.00	0.01	0.00
SL 1 (controls)	0.13	0.13	0.12	0.11	0.11	0.11	0.11	0.12
SE	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Service Learning 2	0.10	0.09	0.08	0.06	0.05	0.05	0.05	0.06
SE	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02
p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SL 2 (controls)	0.13	0.12	0.12	0.12	0.12	0.12	0.12	0.12
SE	0.01	0.02	0.01	0.02	0.02	0.02	0.02	0.02
p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Service Learning 3	0.36	0.36	0.36	0.36	0.25	0.25	0.26	0.21
SE	0.03	0.03	0.03	0.03	0.05	0.05	0.05	0.05
p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SL 3 (controls)	0.29	0.29	0.29	0.33	0.28	0.28	0.28	0.24
SE	0.03	0.03	0.03	0.04	0.05	0.05	0.05	0.05
p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Multiple SL	0.12	0.08	0.09	0.04	0.04	0.04	0.04	0.03
SE	0.02	0.01	0.01	0.02	0.02	0.02	0.02	0.02
p-value	0.00	0.00	0.00	0.02	0.01	0.01	0.01	0.04
Multiple SL (controls)	0.14	0.11	0.12	0.07	0.08	0.08	0.08	0.07
SE	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Appendix 6. Predicted increase in the probability of transfer: Model comparison, weighted and matched samples, 9-term tracking (fall 2007–summer 2020)

Estimand / PSA type	ATE / Inverse Probability of Treatment W.			ATT	ATT / Propensity Score Matching			
PSA model	GBR Modeling	Boosted regres.	Logistic	Kernel	Neighbors: 6	Neighbors: 4	Neighbors: 2	1-to-1 match
Any Service Learning	0.04	0.06	0.01	0.01	0.02	0.02	0.02	0.01
SE	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
p-value	0.00	0.00	0.03	0.08	0.04	0.04	0.04	0.37
Any SL (controls)	0.04	0.05	0.02	0.02	0.02	0.02	0.02	0.01
SE	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
p-value	0.00	0.00	0.01	0.08	0.08	0.06	0.04	0.27
Service Learning 1	0.04	0.05	0.01	0.02	0.01	0.01	0.01	0.02
SE	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02
p-value	0.01	0.00	0.33	0.33	0.52	0.46	0.50	0.21
SL 1 (controls)	0.03	0.05	0.02	0.02	0.02	0.02	0.02	0.03
SE	0.01	0.02	0.01	0.02	0.02	0.02	0.02	0.02
p-value	0.01	0.00	0.19	0.28	0.40	0.45	0.28	0.11
Service Learning 2	0.05	0.05	0.04	0.03	0.03	0.03	0.03	0.04
SE	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
p-value	0.00	0.00	0.00	0.02	0.04	0.02	0.03	0.01
SL 2 (controls)	0.06	0.06	0.06	0.05	0.05	0.05	0.06	0.07
SE	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02
p-value	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
Service Learning 3	0.07	0.07	0.07	0.04	-0.05	-0.03	-0.01	-0.02
SE	0.03	0.03	0.03	0.05	0.05	0.05	0.05	0.05
p-value	0.02	0.02	0.02	0.36	0.37	0.60	0.78	0.62
SL 3 (controls)	0.01	0.01	0.01	-0.02	-0.06	-0.04	-0.02	-0.02
SE	0.03	0.03	0.03	0.05	0.05	0.05	0.05	0.05
p-value	0.69	0.69	0.70	0.67	0.26	0.47	0.75	0.72
Multiple SL	0.03	0.02	0.02	0.01	0.00	0.00	0.00	0.00
SE	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00
p-value	0.05	0.17	0.15	0.68	0.86	0.83	0.84	
Multiple SL (controls)	0.01	0.01	0.01	-0.01	-0.03	-0.03	-0.03	-0.04
SE	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
p-value	0.49	0.41	0.64	0.47	0.19	0.17	0.07	0.06

Appendix 7. OLS estimates of impact of service learning participation on final GPA: Model comparison, weighted and matched samples, fall 2007–summer 2020

Estimand / PSA type	ATE / Inverse Probability of Treatment W.			ATT	ATT / Propensity Score Matching			
PSA model	GBR Modeling	Boosted regres.	Logistic	Kernel	Neighbors: 6	Neighbors: 4	Neighbors: 2	1-to-1 match
Any Service Learning	0.25	0.36	0.13	0.10	0.09	0.09	0.10	0.10
SE	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Any SL (controls)	0.30	0.33	0.22	0.19	0.18	0.18	0.18	0.19
SE	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Service Learning 1	0.22	0.18	0.09	0.08	0.06	0.06	0.07	0.08
SE	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.04
p-value	0.00	0.00	0.00	0.00	0.06	0.05	0.05	0.03
SL 1 (controls)	0.30	0.28	0.25	0.22	0.19	0.19	0.18	0.19
SE	0.03	0.04	0.04	0.03	0.03	0.04	0.04	0.04
p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Service Learning 2	0.27	0.24	0.19	0.12	0.10	0.11	0.11	0.12
SE	0.02	0.03	0.03	0.02	0.03	0.03	0.03	0.03
p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SL 2 (controls)	0.27	0.27	0.23	0.19	0.19	0.20	0.23	0.25
SE	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.04
p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Service Learning 3	0.74	0.74	0.74	0.61	0.28	0.29	0.30	0.30
SE	0.06	0.06	0.06	0.06	0.07	0.07	0.08	0.10
p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SL 3 (controls)	0.57	0.57	0.57	0.48	0.32	0.34	0.34	0.31
SE	0.06	0.06	0.06	0.05	0.06	0.07	0.08	0.09
p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Multiple SL	0.24	0.13	0.15	0.05	0.06	0.07	0.07	0.08
SE	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
p-value	0.00	0.00	0.00	0.08	0.04	0.01	0.03	0.02
Multiple SL (controls)	0.26	0.20	0.21	0.13	0.15	0.17	0.19	0.18
SE	0.04	0.04	0.04	0.03	0.04	0.04	0.04	0.05
p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Appendix 8. Cox proportional hazards model for time to graduation: Model comparison, weighted and matched samples, 9-term tracking (fall 2007–summer 2020)

Estimand / PSA type	ATE / Inverse Probability of Treatment W.			ATT	ATT / Propensity Score Matching			
PSA model	GBR Modeling	Boosted regress.	Logistic	Kernel	Neighbors: 6	Neighbors: 4	Neighbors: 2	1-to-1 match
Any Service Learning	1.12 *	1.15 *	1.11 *	1.10 *	1.10 *	1.09 *	1.10 *	1.10 *
	[1.06, 1.19]	[1.07, 1.24]	[1.05, 1.18]	[1.02, 1.18]	[1.03, 1.18]	[1.02, 1.18]	[1.02, 1.18]	[1.03, 1.19]
Any SL (controls)	1.14 *	1.18 *	1.13 *	1.10 *	1.10 *	1.09	1.09	1.09 *
	[1.06, 1.23]	[1.09, 1.28]	[1.05, 1.21]	[1.01, 1.21]	[1.01, 1.21]	[0.99, 1.19]	[0.99, 1.19]	[1.00, 1.20]
Service Learning 1	0.97	1.01	0.98	1.00	1.01	1.03	1.03	1.09
	[0.88, 1.07]	[0.91, 1.12]	[0.89, 1.08]	[0.87, 1.15]	[0.88, 1.16]	[0.90, 1.18]	[0.90, 1.19]	[0.95, 1.26]
SL 1 (controls)	1.04	1.04	1.06	1.05	1.06	1.09	1.13	1.16
	[0.93, 1.16]	[0.93, 1.17]	[0.95, 1.18]	[0.88, 1.25]	[0.89, 1.26]	[0.92, 1.30]	[0.95, 1.34]	[0.98, 1.38]
Service Learning 2	1.13 *	1.08	1.10 *	1.05	1.03	1.02	1.03	1.00
	[1.04, 1.22]	[0.99, 1.18]	[1.01, 1.19]	[0.94, 1.17]	[0.92, 1.15]	[0.92, 1.14]	[0.92, 1.15]	[0.90, 1.12]
SL 2 (controls)	1.21 *	1.17 *	1.19 *	1.16 *	1.16 *	1.14	1.15	1.12
	[1.09, 1.34]	[1.04, 1.30]	[1.07, 1.31]	[1.01, 1.34]	[1.01, 1.34]	[0.99, 1.32]	[1.00, 1.33]	[0.97, 1.30]
Service Learning 3	1.96 *	1.96 *	1.96 *	1.76 *	1.40 *	1.34	1.32	1.21
	[1.67, 2.31]	[1.67, 2.31]	[1.66, 2.31]	[1.25, 2.49]	[1.04, 1.90]	[0.99, 1.81]	[0.97, 1.78]	[0.90, 1.61]
SL 3 (controls)	1.67 *	1.67 *	1.65 *	1.68 *	1.52 *	1.45 *	1.34	1.23
	[1.41, 1.98]	[1.41, 1.98]	[1.39, 1.96]	[1.13, 2.51]	[1.07, 2.15]	[1.02, 2.07]	[0.93, 1.92]	[0.87, 1.74]
Multiple SL	1.17 *	1.13 *	1.17 *	1.10	1.12	1.14 *	1.15 *	1.09
	[1.06, 1.29]	[1.02, 1.25]	[1.05, 1.29]	[0.97, 1.26]	[0.98, 1.28]	[1.00, 1.30]	[1.01, 1.32]	[0.96, 1.25]
Multiple SL (controls)	1.04	1.02	1.05	0.99	1.01	1.04	1.03	0.99
	[0.92, 1.17]	[0.90, 1.16]	[0.93, 1.19]	[0.83, 1.18]	[0.84, 1.21]	[0.87, 1.25]	[0.86, 1.24]	[0.83, 1.18]

* Indicates statistically significant result (95% confidence interval does not include 1). Confidence intervals in brackets.

PSA: Propensity Score Analysis. ATE: Average Treatment Effect. ATT: Average Treatment Effect for the Treated. GBR: Generalized Boosted Regression.

Appendix 9. Cox proportional hazards model for time to transfer: Model comparison, weighted and matched samples, 9-term tracking (fall 2007–summer 2020)

Estimand / PSA type	ATE / Inverse Probability of Treatment W.			ATT	ATT / Propensity Score Matching			
PSA model	GBR Modeling	Boosted regress.	Logistic	Kernel	Neighbors: 6	Neighbors: 4	Neighbors: 2	1-to-1 match
Any Service Learning	0.84 *	0.75 *	0.85 *	0.89 *	0.90 *	0.89 *	0.89 *	0.86 *
	[0.78, 0.89]	[0.69, 0.83]	[0.80, 0.91]	[0.82, 0.97]	[0.83, 0.97]	[0.82, 0.97]	[0.82, 0.97]	[0.80, 0.94]
Any SL (controls)	0.76 *	0.77 *	0.76 *	0.80 *	0.80 *	0.79 *	0.80 *	0.77 *
	[0.70, 0.83]	[0.70, 0.85]	[0.70, 0.82]	[0.72, 0.88]	[0.72, 0.88]	[0.71, 0.88]	[0.72, 0.89]	[0.69, 0.85]
Service Learning 1	0.84 *	0.93	0.84 *	0.92	0.91	0.91	0.91	0.92
	[0.75, 0.95]	[0.83, 1.05]	[0.75, 0.94]	[0.78, 1.08]	[0.78, 1.07]	[0.78, 1.07]	[0.77, 1.07]	[0.78, 1.08]
SL 1 (controls)	0.77 *	0.84 *	0.75 *	0.80 *	0.81 *	0.81 *	0.84	0.89
	[0.67, 0.87]	[0.73, 0.96]	[0.66, 0.86]	[0.66, 0.98]	[0.67, 0.99]	[0.66, 0.98]	[0.69, 1.02]	[0.73, 1.08]
Service Learning 2	0.95	0.94	0.97	0.95	0.93	0.94	0.94	0.96
	[0.87, 1.05]	[0.85, 1.04]	[0.88, 1.06]	[0.83, 1.08]	[0.82, 1.06]	[0.83, 1.07]	[0.83, 1.07]	[0.84, 1.09]
SL 2 (controls)	0.93	0.93	0.94	0.95	0.92	0.92	0.97	1.00
	[0.84, 1.04]	[0.83, 1.05]	[0.85, 1.05]	[0.80, 1.12]	[0.78, 1.08]	[0.78, 1.09]	[0.83, 1.15]	[0.84, 1.17]
Service Learning 3	0.77	0.77	0.77	0.70	0.63 *	0.65 *	0.66 *	0.68
	[0.58, 1.01]	[0.58, 1.01]	[0.58, 1.01]	[0.45, 1.10]	[0.42, 0.95]	[0.43, 0.98]	[0.43, 1.00]	[0.45, 1.03]
SL 3 (controls)	0.67 *	0.67 *	0.67 *	0.59 *	0.59 *	0.59 *	0.62	0.65
	[0.51, 0.87]	[0.51, 0.87]	[0.51, 0.87]	[0.35, 1.00]	[0.36, 0.95]	[0.36, 0.97]	[0.37, 1.03]	[0.39, 1.10]
Multiple SL	0.84 *	0.87 *	0.88 *	0.95	0.92	0.92	0.91	0.91
	[0.74, 0.96]	[0.77, 0.98]	[0.77, 0.99]	[0.82, 1.11]	[0.79, 1.07]	[0.79, 1.08]	[0.78, 1.06]	[0.79, 1.06]
Multiple SL (controls)	0.70 *	0.75 *	0.72 *	0.72 *	0.68 *	0.68 *	0.65 *	0.67 *
	[0.60, 0.80]	[0.65, 0.86]	[0.63, 0.83]	[0.59, 0.88]	[0.56, 0.83]	[0.56, .83]	[0.53, 0.79]	[0.55, 0.81]

* Indicates statistically significant result (95% confidence interval does not include 1). Confidence intervals in brackets.

PSA: Propensity Score Analysis. ATE: Average Treatment Effect. ATT: Average Treatment Effect for the Treated. GBR: Generalized Boosted Regression.

References

Allison, P. D. (1984). *Event history analysis: Regression for longitudinal event data*. Sage Publications.

Anderson, K. L., Boyd, M., Marin, K. A., & McNamara, K. (2019). Reimagining service-learning: Deepening the impact of this high-impact practice. *Journal of Experiential Education, 42*(3), 1-20. https://doi.org/10.1177%2F1053825919837735

Andrade, M. S. (2007). Learning communities: Examining positive outcomes. *Journal of College Student Retention: Research, Theory, & Practice*, *9*(1), 1–20. https://doi.org/10.2190%2FE132-5X73-681Q-K188

Angrist, J. D. (1997). Conditional independence in sample selection models. *Economics Letters*, *54*(2), 103–112. https://doi.org/10.1016/S0165-1765(97)00022-0

Association of American Colleges & Universities (AAC&U). (2007). *College learning for the new global century: A report from the National Leadership Council for Liberal Education & American's Promise.* The author.

Astin, A. W., Vogelgesang, L. J., Ikeda, E. K., & Yee, J. A. (2000). How service learning affects students. *Higher Education, 144*.

Austin, P. C. (2011a). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46*(3), 399-424. https://doi.org/10.1080/00273171.2011.568786

Austin, P. C. (2011b). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics, 10*(2), 150-161. https://doi.org/10.1002/pst.433

Austin, P. C. (2016). Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Statistics in Medicine, 35*(3), 5642–55. https://doi.org/10.1002/sim.7084

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine, 34*, 3661-3679. https://doi.org/10.1002/sim.6607

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, *61*(4), 962–973. https://doi.org/10.1111/j.1541-0420.2005.00377.x

Bennett, D. S. (1999). Parametric models, duration dependence, and time-varying data revisited. *American Journal of Political Science, 43*(1), 256-270. https://doi.org/10.2307/2991793

Box-Steffensmeier, J. M., & Jones, B. S. (2004). *Event history modeling: A guide for social scientists*. Cambridge University Press.

Bringle, R., Hatcher, J., & McIntosh, R. (2006). Analyzing Morton's typology of service paradigms and integrity. *Michigan Journal of Community Service Learning, 13*, 5-15. https://www.elon.edu/u/service-learning/wp-content/uploads/sites/519/2018/07/46963247.pdf

Brownell, J. E., & Swaner, L. E. (2010). *Five high-impact practices: Research on learning outcomes, completion, and quality*. Association of American Colleges and Universities.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of the propensity score matching. *Journal of Economic Surveys*, *22*(1), 31-72.

Cefalu, M., & Buenaventura, M. (2017). *Propensity scores for multiple treatments: A tutorial on the MNPS command for Stata users*. RAND Corporation.

Cefalu, M., Liu, S., & Marti, C. (2015). *Toolkit for weighting and analysis of nonequivalent groups: A tutorial on the Twang commands for Stata users*. RAND Corporation.

Center for Community College Student Engagement (CCCSE). (2012). *A matter of degrees: Promising practices for community college student success (a first look)*. The University of Texas at Austin.

Cleves, M., Gould, W. W., & Marchenko, Y. V. (2016). *An introduction to survival analysis using Stata* (3rd ed., revised). Stata Press.

Cole, S. R., & Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology, 168*(6), 656-664. https://doi.org/10.1093/aje/kwn164

Darche, S., & Arnold, M. P. (2004). *The benefits of work-based learning and occupational coursework in the California community colleges* (Report). Hatchuel Tabernik and Associates. https://www.mendocino.edu/sites/default/files/docs/work-experience/Benefits%20of%20WBL%20Report%205-04.pdf

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association, 94*(448), 1053-1062.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guided to boosted regression trees. *Journal of Animal Ecology, 77*(4), 802-813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Engberg, M. E., & Fox, K. (2011). Exploring the relationship between undergraduate service-learning experiences and global perspective-taking. *Journal of Student Affairs Research and Practice, 48*(1), 85–105. http://dx.doi.org/10.2202/1949-6605.6192

Eyler, J., & Giles, D. (2001). *At a glance: What we know about the effects of service-learning on college students, faculty, institutions, and communities, 1993–2000* (3rd. ed.). Corporation for National and Community Service.

Eyler, J., Giles, D., & Braxton, J. (1997). The impact of service-learning on college students. *Michigan Journal of Community Service Learning, 4*, 5-15.

Felten, P., & Clayton, P. H. (2011). Service-learning. *New Directions for Service & Learning, 128,* 75-84. https://doi.org/10.1002/tl.470

Feng, P., Zhou, X. , Zou, Q., Fan, M., & Li, X. (2012). Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in Medicine*, *31*(7), 681-97. https://doi.org/10.1002/sim.4168

Finley, A. (2012). *Making progress? What we know about the achievement of liberal education learning outcomes*. Association of American Colleges and Universities.

Finley, A., & McNair, T. (2013). *Assessing underserved students engagement in high-impact practices*. Association of American Colleges and Universities.

Gallini, S. M., & Moely, B. E. (2003). Service-learning and engagement, academic challenge, and retention. *Michigan Journal of Community Service Learning, 10*(1), 5–14.

Garrido, M. M., Kelley, A. S., Paris, J., Roza, K., Meier, D. E., Morrison, R. S., & Aldridge, M. D. (2014). Methods for constructing and assessing propensity scores. *Health Services Research, 49*(5), 1701-1720. https://doi.org/10.1111/1475-6773.12182

Goff, J., Hill, E., Eckhoff, A., & Dice, T. (2020). Examining the high-impact practice of service-learning: Written reflections of undergraduate recreation majors. *SCHOLE: A Journal of Leisure Studies and Recreation Education, 35*(1), 1-13. https://doi.org/10.1080/1937156X.2020.1720444

Guo, S., & Fraser, M. W. (2015). *Advanced quantitative techniques in the social sciences: Vol. 11. Propensity score analysis: Statistical methods and applications* (2nd ed). SAGE Publications, Inc.

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*, *15*(3), 234–249. https://doi.org/10.1037/a0019623

Hatch, D. K. (2013). *Student engagement and the design of high-impact practices at community colleges* (Unpublished doctoral dissertation). The University of Texas, Austin.

Heckman, J. J., Ichimura, H, & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies, 64*(4), 605-654. https://doi.org/10.2307/2971733

Heckman, J. J., Ichimura, H, & Todd, P. E. (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies, 65*, 261-94.

Heckman, J. J., & Robb, R. (1985). Alternative methods for estimating the impact of interventions. In J. Heckman & B. Singer (Eds.), *Longitudinal analysis of labor market data* (pp. 156-245). Cambridge University Press.

Hirano, K., & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, *2*, 259-278. https://doi.org/10.1023/A:1020371312283

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica, 71*(4), 1161–89.

Hu, S., & McCormick, A. C. (2012). An engagement-based student typology and its relationship to college outcomes. *Research in Higher Education, 53*, 738–754. https://doi.org/10.1007/s11162-012-9254-7

Imai, K., & van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*; *99*(467), 854–866. https://doi.org/10.1198/016214504000001187

Imbens, G. W. (2000). The role of propensity score in estimating dose-response functions. *Biometrika, 87*(3), 706-710. https://doi.org/10.1093/biomet/87.3.706

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, *86*(1), 4-29.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments of the econometrics of program evaluation. *Journal of Economic Literature, 47*(1), 5-86.

Inkelas, K. K., Vogt, K. E., Longerbeam, S. D., Owen, J., & Johnson, D. (2006). Measuring outcomes of living-learning programs: Examining college environments and student learning and development. *The Journal of General Education, 55*(1), 40–76. https://doi.org/10.1353/jge.2006.0017

Jann, B. (2017). *Why propensity scores should be used for matching*. Presentation at the German Stata Users' Group Meetings.

Joffe, M. M., Ten Have, T. R., Feldman, H. I., & Kimmel, S. E. (2004). Model selection, confounder control, and marginal structural models: Review and new applications. *The American Statistician, 58*(4), 272–279. https://doi.org/10.1198/000313004X5824

Johnson, S. R., & Stage, F. K. (2018). Academic engagement and student success: Do high-impact practices mean higher graduation rates? *The Journal of Higher Education, 89*(5), 753-781. https://doi.org/10.1080/00221546.2018.1441107

Kraft, R. J. (1996). Service learning: An introduction to its theory, practice, and effects. *Education and Urban Society, 28*(2), 131–59.

Kilgo, C. A., Ezell Sheets, J. K., & Pascarella, E. T. (2013). *Do high-impact practices actually have high impact on student learning? Some initial findings*. Paper presented at the Annual Conference of the Association for the Study of Higher Education, St. Louis, MO. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.697.6006&rep=rep1&type=pdf

Kilgo, C. A., Ezell Sheets, J. K., & Pascarella, E. T. (2015). The link between high-impact practices and student learning: Some longitudinal evidence. *Higher Education, 69*, 509–525. https://doi.org/10.1007/s10734-014-9788-z

King, G. (2015, September 11). *Why propensity scores should not be used for matching* [Video]. YouTube. https://www.youtube.com/watch?v=rBv39pK1iEs

King, G., & Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis, 27*(4), 435-454. https://doi.org/10.1017/pan.2019.11

King, G., Nielsen, R., Coberley, C., Pope, J. E., & Wells, A. (2011). *Comparative effectiveness of matching methods for causal inference*. https://gking.harvard.edu/publications/comparative-effectiveness-matching-methods-causal-inference

Kuh, G. (2008). *High-impact educational practices: What they are, who has access to them, and why they matter*. Association of American Colleges and Universities.

Kuh, G., & O'Donnell, L. (2013). *Ensuring quality and taking high-impact practices to scale*. Association of American Colleges and Universities.

Largent, L., & Horinek, J. B. (2008). Community colleges and adult service-learners: Evaluating a first-year program to improve implementation. *New Directions for Adult and Continuing Education, 118*, 37–47. doi:10.1002/ace.294

Lee, B. K., Lessler, J., Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine, 29*(3), 337-46. https://doi.org/10.1002/sim.3782

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., & Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine, 32*(19), 3388-3414. https://doi.org/10.1002/sim.5753

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted-regression for evaluating causal effects in observational studies. *Psychological Methods, 9*(4), 403-425. https://doi.org/10.1037/1082-989x.9.4.403

Morgan, S. L., & Harding, D. J. (2006). Matching estimators of causal effects: Prospects and pitfalls in theory and practice. *Sociological Methods & Research*, *35*(1), 3-60.

Morgan, S. L., & Todd, J. J. (2008). A diagnostic routine for the detection of consequential heterogeneity of causal effects. *Sociological Methodology*, *38*(1), 231–281.

Murnane, R. J., & Willet, J. B. (2011). *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press.

Pascarella, E. G., & Terenzini, P. T. (2005). *How college affects students: A third decade of research* (Vol. 2). Jossey-Bass.

Prentice, M., & Robison, G. (2010). *Improving student learning outcomes with service learning* (Report AACC-RB-10-1). American Association of Community Colleges.

Provencher, A., & Kassel, R. (2017). High-impact practices and sophomore retention: Examining the effects of selection bias. *Journal of College Student Retention: Research, Theory & Practice, 21*(2), https://doi.org/10.1177%2F1521025117697728

RAND. (n.d.). *A tutorial on the TWANG commands for Stata users*. RAND Corporation.

Ridgeway, G. (2007, August 3*). Generalized boosted models: A guide to the GBM package*. http://www.saedsayad.com/docs/gbm2.pdf

Ridgeway, G., McCaffrey, D., Morral, A., Burgette, L., & Griffin, B. A. (2020, February 26). *Toolkit for weighting and analysis of nonequivalent groups: A tutorial for the twang package*. RAND Corporation.

Rosenbaum, P. R., & Rubin D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*(387), 516–524.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688–701. https://doi.org/10.1037/H0037350

Rubin D. B., & Thomas N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics*, *52*(1), 249-264. https://doi.org/10.2307/2533160

Sandeen, C. (2012). High-impact educational practices: What we can learn from the traditional undergraduate setting. *Continuing Higher Education Review*, *76*, 81-89. https://files.eric.ed.gov/fulltext/EJ1000654.pdf

Sato, T., & Matsuyama, Y. (2003). Marginal structural models as a tool for standardization. *Epidemiology, 14*(6), 680-686. https://doi.org/10.1097/01.ede.0000081989.82616.7d

Schonlau, M. (2005). Boosted regression (boosting): An introductory tutorial and a Stata plugin. *The Stata Journal*, *5*(3), 330-354.

Seifert, T. A., Gillig, B., Hanson, J. M., Pascarella, E. T., & Blaich, C. F. (2014). The conditional nature of high impact/good practices on student learning outcomes. *The Journal of Higher Education, 85*(4), 531-564. https://doi.org/10.1080/00221546.2014.11777339

Sigmon, R. (1979). Service-learning: Three principles. *Synergist*, *8*(10), 9–11.

Simons, L., & Cleary, B. (2006). The influence of service learning on students’ personal and social development. *College Teaching, 54*(4), 307–319. https://doi.org/10.3200/CTCH.54.4.307-319

Smith, J. A., & Todd, P. E. (2001). Reconciling conflicting evidence on the performance of propensity score methods. *American Economic Review, 91*(2), 112–118.

Swaner, L. E., & Brownell, J. E. (2009). *Outcomes of high impact practices for underserved students: A review of the literature*. Prepared for the Association of American Colleges and Universities (AAC&U) Project USA. Retrieved from http://www.aacu.org/inclusive_excellence/documents/DRAFTProjectUSALiteratureReview.pdf

Tennessee Board of Regents (TBR). (n.d.). *HIP taxonomy: Service learning*. https://www.tbr.edu/academics/studentaffairs/hip-taxonomy-service-learning

Tennessee Board of Regents (TBR). (2019, January). *A vision for high impact practices in Tennessee*. Presentation at the HIP statewide conference, Nashville, TN. https://www.tbr.edu/sites/default/files/media/2019/07/HighImpactPractices_201901.pdf

Tennessee Board of Regents (TBR). (2020, June). *Student success and the first-year experience: Early evidence from TBR community colleges* (Working paper). The author. https://www.tbr.edu/sites/default/files/media/2020/07/FYSResearchSummaryApril2020.pdf

Tennessee Board of Regents (TBR). (2021, February). *Student engagement and college outcomes: Analysis of CCSSE data on Tennessee community college students* (Working paper). The author. https://www.tbr.edu/policy-strategy/presentations-and-papers

Thoemmes. F., & Ong, A. D. (2015). A primer on inverse probability of treatment weighting and marginal structural models. *Emerging Adulthood, 4*(1), 40-59. https://doi.org/10.1177/2167696815621645

Valentine, J., & Price, D. (2021). *Scaling high impact practices to improve community college students outcomes: Evidence from the Tennessee Board of Regents* (Policy Brief). Lumina Foundation, DVP-Praxis.

Valentine, J., Price, D., & Yang, H. (2021). *High-impact practices and gains in student learning: Evidence from Georgia, Montana, and Wisconsin* (Issue paper). Lumina Foundation.

Vogt, W. P. (2005). *Dictionary of statistics and methodology: A nontechnical guide for the social sciences* (3rd. ed). SAGE Publications, Inc.

Wang. Y., Xie, J., Liu, X., Du, J., Wu, M., Huang, W., & Deng, D. (2019). Logistic regression and generalized boosted modeling in inverse probability of treatment weighting: A simulation and case study of outpatients with coronary heart disease. *Journal of Epidemiology and Public Health Reviews, 4*(3), 1-9. dx.doi. org/10.16966/2471-8211.178

Watson, C. E., Kuh, G. D., Rhodes, T., Light, T. P., & Chen, H. L. (2016). ePortfolios-The eleventh high impact practice. *International Journal, 6*(2), 65–69. http://www.theijep.com/pdf/IJEP254.pdf

Wolniak, G., & Engberg, M. (2015, April 18). *The influence of 'high-impact' college experiences on early career outcomes*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.