TIAA Institute

Research Dialogue | Issue no. 215
August 2024

Counterfactual predictive analytics for enhancing STEM student success: Using the community college sector as an entry point

Abstract

This study offers a framework for affordable STEM human resource formation using the community college sector as an entry point. Using predictive analytics and counterfactual evidence, we identify first-year community college students with high probabilities of four-year graduation in STEM disciplines with two most recent national samples: National Educational Longitudinal Study (NELS), 1988–2000, and Education Longitudinal Study (ELS), 2002–2012. Findings revealed that while the prevalence of these counterfactual community college students doubled in the most recent sample (11.7% in NELS versus 22.7% in ELS), they also became less likely to attain a four-year degree in STEM. In the NELS sample the probability of a BS in STEM degree attainment was 18.2% whereas in the ELS sample it became 14%. At the cohort level, we observed a 70% decrease in the likelihood of counterfactual students attaining a four-year STEM degree without intervention. Socioeconomic status as the single most important driver of STEM attainment individual attributes were less predictive in the ELS sample. This study lays the groundwork for future interventions to improve success rates among community college students in STEM fields, emphasizing the importance of early identification and tailored support. Future studies can build upon our counterfactual framework to identify these students and enhance prospects for STEM degree attainment.

Manuel S. González Canché
University of Pennsylvania
TIAA Institute Fellow

Chelsea Zhang
University of Pennsylvania

Introduction

In their very inception in the early 1900s, public two-year community colleges were originally conceived to offer programs extending from high schools as the 13th and 14th years of formal education (Clark, 1960; Cohen, 1987; Garrison, 1975; Helland, 1987; Vaughan, 1995). However, by the 1920s, community college course offerings began mirroring the first two years of coursework available at four-year institutions (Clark, 1960; Spindt, 1957), allowing students to pursue education that enabled them to transfer to a four-year institution starting in their "junior" year of college (Spindt, 1957).

Fast forward more than a century, community colleges—or public two-year colleges (used interchangeably henceforth)—have evolved to encompass two defining roles. First, they serve as steppingstones to four-year degrees (González Canché, 2018). Secondly, they function as a vital and affordable entry point to higher education for minoritized, low-income, and first-generation-in-college students (González Canché, 2014, 2018, 2017b, 2020, 2022). The amalgamation of these roles means that, for at least the past three decades, the majority of these students have aimed to transfer to a four-year institution. Analyzing two nationally representative samples of students transitioning from high school to college (National Education Longitudinal Study of 1988 and Education Longitudinal Study of 2002), González Canché (2020) found that in the 1990s and 2000s, 69% and 71% of high school seniors, respectively, who began college in the community college sector, expected to attain at least a bachelor's degree. Similarly, estimates of another recent nationally representative study on high school to college transitions (High School Longitudinal Study—HSLS, 2009–ongoing) showed that, in the 2010s, 60% of community college entrants expected to attain at least a bachelor's degree (authors' estimates based on restricted-use HSLS data). These nationally representative estimates indicate that, across three decades, about two-thirds of community college students have expected to attain a four-year degree, thus validating their perceived role as gateways to higher levels of education.

However, despite these elevated expectations, empirical evidence consistently reveals that less than a third of community college students with aspirations for a four-year degree actually transfer to a four-year college, resulting in a "cohort bachelor's completion rate"[1] of 14.2% (Shapiro et al., 2020). This persistent gap, referred to as the *baccalaureate gap* (Dougherty, 1992), has been substantiated across four decades of research, which indicates a consistent average reduction of 28.4% (standard deviation = 19.2%) in community college students' probabilities of attaining a bachelor's degree compared with their four-year counterparts (González Canché, 2018; Figure 2 https://cutt.ly/VQXGWf3).

Notably, when researchers focus on two-year entrants who transferred to a four-year institution (referred to as raising juniors), around 46% of these transfer students attain a four-year degree within six years of their initial college enrollment (Shapiro et al., 2020). This figure is important because studies in this area have concluded that when community college students do transfer to a four-year institution, they are as likely to attain a four-year degree as their counterparts who started college in the four-year sector (Dietrich & Lichtenberger, 2015; Lichtenberger & Dietrich, 2017; Monaghan & Attewell, 2015).

Together, these findings carry significant implications. First, despite enrolling a financially constrained and minoritized student segment, these comparable success rates of transfers and four-year natives mean that the two-year path to a four-year degree is feasible for a subset of two-year entrants. Second, when considering that at least about two-thirds of community college students aspire to transfer, but only a small portion of them achieve doing so, then we could conclude that if the transfer process is streamlined, we could potentially also increase the proportion of community college students attaining four-year degrees. Finally, if we consider that the 2022 national six-year completion rate among public and private not-for-profit four-year students is 73.65% (National Student Clearinghouse, 2022),[2] it could be argued that if we device a plan to identify two-year entrants with similar academic performance as their four-year counterparts, such two-year entrants should have about a 75% chance of attaining a four-year degree as well.

Another noteworthy aspect of relying on the two-year path toward a four-year degree is its enhanced affordability. Studies on the financial impact of this path—particularly in terms of student loan debt reliance—consistently indicate a reduction of about 10 percentage points in total debt accumulation (González Canché, 2014, 2020; Hu, et al., 2017). Despite longer enrollment times and heightened opportunity costs (i.e., loss of revenue associated with full-time enrollment), this path may still present a more affordable option for a subset of community college students

1 Ratio of the number of community college entrants in year X who attained a four-year degree within six years divided by the total number of community college entrants in year X (Shapiro et al., 2020).

2 This graduation rate is 69% for public four-year colleges and 78.3% for private not-for-profit colleges. Our 73.65 percent estimate was calculated by adding these percentages and dividing them by 2.

(González Canché, 2020). Together, these degree attainment and financial outcomes are appealing for they underscore the significance of studying innovative strategies to strengthen the two-year path to a four-year degree of *minoritized, low-income students,* presenting meaningful implications for closing educational gaps (González Canché, 2017a, 2022; González Canché, et al., 2023).

Community college students and STEM degree attainment

So far, our discussion has centered on overall degree attainment, with no discussion of the possibility or feasibility of using the two- to four-year path toward a STEM degree. Given that "STEM workers with a bachelor's degree or higher have a median salary that is 47% greater than that of non-STEM workers with a bachelor's degree or higher" (National Science Board, 2021), if two-year students attain a four-year degree in STEM, this will result in significant social and economic upward mobility. This prospect is particularly pertinent for community college students from minoritized and low-income backgrounds. Accordingly, in the following paragraphs we delve into existing research on this two-year STEM topic.

Malcom and Feder (2016) edited a National Academies of Sciences, Engineering, Medicine report on the barriers and opportunities for two- and four-year STEM degrees. In their report they inquired about another type of STEM degree departure: STEM drop-out, which involves not completing a degree in STEM, either by moving to a non-STEM degree or leaving before completing any degree. The report fundamentally sought to identify strategies to increase STEM degree completion while acknowledging that students' pathways to a four-year STEM degree do not necessarily begin in a four-year institution. Investigating this less conventional, more indirect pathway, González Canché (2017b) identified that approximately 10% of STEM PhD holders in the National Science Foundation's Survey of Doctorate Recipients (2021) reported having started college in the two-year sector. Similar to the preceding discussion on comparable success rates of transfers and four-year students, the presence of these "community college scientists" (a term that signals their initial college formation in the public two-year sector—González Canché, 2017b) also means that these two-to-four-year paths have translated into the attainment of the highest and most prestigious terminal degree in the United States (González Canché, 2017b).

Although this line of study is scarce, authors (González Canché, 2017b; Malcom & Feder, 2016) have found that a subset of community college entrants not only successfully transferred to a four-year institution, but also graduated with a STEM degree. Revisiting the transfer-out estimates, let us note that about two-thirds of community college students with a four-year degree aspiration didn't transfer to a four-year institution (Shapiro, et al., 2020). Since this latter group represents the majority of community college entrants with four-year degree aspirations, in this study we pose this guiding question: *Can we identify initial community college students with high probabilities of four-year graduation in STEM disciplines?*

Note that this question is not restricted to transfer, but rather it goes to actual four-year degree graduation probabilities in STEM. Relatedly, this question is grounded in the assumption that there exists a subset of community college students with high probabilities of attaining a four-year STEM degree. If this is true, so far, we have no idea of the prevalence of these cases with high success probabilities, and, consequently, we also don't know the degree to which they're realizing their potential for such a degree. As stated above, STEM degrees have significant salary-related payoffs at the individual level, which is important for social and economic mobility. Moreover, from a societal perspective, as "science and technology development tighten across the world, the production of STEM graduates is of increasing importance for any country aspiring to remain competitive in scientific production worldwide" (González Canché, 2017b). From these individual and communal perspectives—and underscoring the important goal of increasing STEM production—it's worth assessing the extent to which the community college sector's role in serving as a steppingstone to a STEM four-year degree may be strengthened. Accordingly, our study aims to develop a strategy to identify community college entrants with high propensities of completing a four-year degree in a STEM field or discipline.

Purpose and practical relevance

Community colleges serve as important and affordable entry points to higher education for minoritized, low-income, and first-generation students. Yet two-year students experience lower academic success rates than their four-year counterparts. This study aims to develop, operationalize, and test an identification framework to address three key objectives: (a) to illuminate the prevalence of community college students with high probabilities of attaining STEM four-year degrees using data from the two most recent national studies, Education Longitudinal Study (2002–2012) and the National Educational Longitudinal Study (1988–2000), (b) to explore differences between observed and expected outcomes of these students, and (c) use machine learning to identify predictors of success.

This study introduces an identification framework that detects community college entrants with high probabilities of

attaining STEM four-year degrees, an approach that hasn't been applied previously to this line of inquiry. By identifying community college entrants with high potential for STEM degree attainment, this framework facilitates the assessment of unrealized contributions from minoritized students in STEM fields, potentially justifying the urgency of developing targeted interventions building from this identification framework.

Research questions

With this discussion in mind, this study addresses the following research questions:

1. Can we develop an innovative method using counterfactual causality and predictive analytics to identify two-year entrants likely to attain a four-year STEM degree (referred as "counterfactual students")?
2. How common are these counterfactual cases among recent nationally representative samples of high school graduates transitioning to college in the United States?
3. What is the distribution of four-year STEM degree attainment among two-year entrants, distinguishing between those identified as counterfactuals and non-counterfactuals? How do these distributions change when considering less than four-year degrees in STEM by these comparison groups? How do these distributions vary across decades?
4. How do the observed outcomes of community college students differ from their expected probabilities of attaining a four-year degree in STEM, and attaining a less than four-year degree in STEM? How do these success rates differ between counterfactual and non-counterfactual community college entrants?
5. What are the best predictors of four-year STEM degree attainment for counterfactual two-year students? How does predictive power change when excluding socioeconomic status to test for sensitivity[3] of these results as well as excluding academic indicators used to identify counterfactual cases? Additionally, how do these results compare across decades?

How have matching methods been used in higher education research?

This section provides a methodologically focused literature review to elucidate the relevance of matching methods for our identification strategy. Emphasis is placed on comparing probabilistic matching with propensity score methods. Reynolds and DesJardins (2009) illustrated the efficacy of matching methods in addressing inferential challenges encountered in nonexperimental education data. Similar to this project, they examined the impact of attending a two-year college on educational attainment. The authors argue that semi-parametric methods, like matching, offer greater flexibility and require fewer assumptions than ordinary least squares (OLS) regression, leading to more accurate or less biased estimates compared to ignoring potential systematic differences across comparison groups. In the context of our participants, ample research (González Canché, 2014; Melguizo et al., 2011; Reynolds & DesJardins, 2009) has shown that two- and four-year students have substantially different distributions in their observed characteristics. Thus, in the absence of a statistical correction, the results would not capture the impact of type of college attended, but the systematic disparities across participants, which incidentally played a role in their ultimate decisions to attach different college types.

To address the challenge that self-selection based on observables poses in this line of study, Reynolds and DesJardins (2009) matched students based on the likelihood of receiving treatment, known as the "propensity score," determined through logistic regression. A propensity score, therefore, simplifies the matching problem to a single dimension while aiming to balance between observed covariates across treated and control groups (Rosenbaum & Rubin, 1983). However, the way such a single score is produced also speaks to the limitations of propensity score matching (PSM). Critiques include that PSM reduces multiple covariates into a single score, and that PSM can still be sensitive to the curse of dimensionality, especially when dealing with a large number of covariates. As the dimensionality increases, finding comparable matches becomes more difficult, potentially resulting in poor matches and biased estimates (Caliendo & Kopeinig, 2008).

Probabilistic matching, as an alternative to PSM, can mitigate some of the disadvantages associated with PSM. The probabilistic matching approach followed mathematical principles outlined by Felligi and Sunter in their 1969 paper, "A Theory of Record Linkage." Record linkage is a data integration technique to identify and link related records across different datasets. The primary goal is to recognize and merge records that correspond to the "same entity,"

3 Sensitivity will be tested to address the fifth research question by intentionally excluding socioeconomic status (and then excluding academic indicators) from the machine learning classification model in order to assess how remaining indicators may vary in their predictive power or relative influence in machine learning specifications—see generalized boosted models (GBM) in the methods section.

even if these records have variations or errors in the data (i.e., spelling errors or similar human mistakes). This process involves comparing specific fields or attributes within records, such as names, addresses, and quantitative indicators and identifiers, to create a unified data set (Dusetzina et al., 2014). One application of record linkage can be found in Osborne and McLaurin's (2006) study, which used probability matching and record linkage to estimate the proportion of former further education (FE) students (comparable to students from community colleges) transferred to Scottish higher education institutions (HEIs). With unadjusted data from the Scottish Further and Higher Education Funding Council (1994–1999) and Higher Education Statistics Agency (1999–2000), they identified comparable entries (first initial, second initial, date of birth, and education council ID) to ascertain whether individuals in four-year universities had a background in further education.

Probabilistic matching, a record linkage technique, is also generally more adept at handling high-dimensional data (Enamorado et al., 2019). It can manage many covariates without explicitly reducing them to a single dimension. Moreover, probabilistic matching can be more robust in dealing with missing data (Sayers et al., 2016) for records may also be linked based on their common missingness—an approach we implemented in this study.

Procedurally, probabilistic matching assesses the likelihood that two records match based on whether they "agree" or "disagree" with the specified identifiers. It is this probabilistic matching via record linkage that we consider essential to our counterfactual identification strategy. Probabilistic matching quantifies the uncertainty inherent in merging procedures (Enamorado et al., 2019); in our study, such uncertainty can be understood as the differences between an individual and his or her matched counterpart. We'll explain further in the Methodology section.

An important concept in probabilistic matching is *blocking.* Blocking reduces the number of data pairs by focusing on specific agreement patterns and may be an integral aspect of the probabilistic matching theory (Felligi & Sunter, 1969). Instead of comparing all possible pairs, which can be computationally intensive, blocking involves restricting comparisons to records with equal values for certain attributes—like attending the same schools or living in the same zip code tabulated areas—known as blocking fields (Sariyar & Borg, 2010). Enamorado et al. (2019) indicated that blocking is helpful in addressing the problem of a low degree of overlap. Although important, we refrained from using blocking because this would naturally limit our possibilities to finding counterfactuals across the full spectrum of community college entrants and bachelor's degree holders in STEM. We are, instead, interested in identifying counterfactual cases regardless of whether they lived in the same state or attended the same high schools.

Methodology

The proposed identification strategy is grounded in the assumption that the detection of community college students with high probabilities of attaining a four-year STEM degree can be accomplished through pre-college data indicators. This identification strategy builds on the counterfactual (Lewis, 1974) and potential outcomes (Rubin, 2005) frameworks. Counterfactual identification involves detecting participants whose "worlds" closely resemble each other. In essence, two individuals are considered good counterfactuals when they share similarities in their living environments, households, monetary and nonmonetary support systems, and academic achievement. Identifying optimal counterfactuals allows for the comparison of potential and actual outcomes. The process begins with identifying (a) a donor pool of the successful cases of interest—defined in this project as four-year degree holders in a STEM discipline—and (b) a pool of two-year students from which feasible counterfactuals will be identified (see Panels (a) and (b) in Figure 1).

Using these pools, the strategy identifies the best counterfactuals of successful cases among the set of two-year students, considering the degree of resemblance in observable characteristics. This resemblance is quantified as a feasibility weight (Sariyar & Borg, 2010), measured on a scale from 0 to 1, where 1 indicates perfect counterfactual identification. As the donors' and two-year students' attributes depart from exact matches, the counterfactual probability approaches zero (see Dusetzina, et al., 2014; Sariyar & Borg, 2010). A representation of this process can be seen in the green column in Panel (d) of Figure 1. Each community college student will have many matches, each assigned a feasibility probability when there are successful cases in the donor pool (Panel (a) of Figure 1). Notably, these counterfactual probabilities don't necessarily span the 0-to-1 range uniformly; some students may have the feasibility weights close to 0, signifying a lack of feasible counterfactuals, whereas others may approach 1 due to the abundance of feasible matches from the donor pool. In this project, the initial identification involves detecting students with at least one minimum feasibility weight of 0.75 and retaining only the maximum value. A feasibility of 0.75 means that the community college student started college with a 75% chance of completing a four-year STEM degree. Pragmatically, this threshold matches the 2022 national six-year completion rate among public and private not-for-profit four-year students reported by the National Student Clearinghouse (2022).

Let us note that without blocking, each community college student will have as many feasibility weights as donor units in the sample. From this perspective, an empirical challenge consisted of devising a plan to retain only the maximum

value after this 0.75 threshold was met. That is, assume a community college student is identified counterfactually with two four-year STEM graduates with feasibility values of 0.85 and 0.92. After applying the identification function, only the 0.92 weight will be retained because it is the maximum value.

As briefly discussed above, additional criteria such as blocking (Dusetzina et al., 2014; Sariyar & Borg, 2010) may be incorporated during feasibility weights detection. For example, matches may be constrained to happen only among low-income status and ethnicity, and then further refining the identification to retain only the maximum feasibility probability or weight after applying these blocking constraints. Although future studies may rely on these blocking mechanisms, we refrained from applying this procedure to expand the possibility of identifying as many two-year students with high probabilities of four-year degree attainment in STEM as possible.

Statistical description

The record linkage theory is a methodology to establish the likelihood of matching components across records by creating a random vector of log odds. This vector is constructed by assigning one score or one weight to each component based on the probability of a match with another component. The process involves calculating log odds ratios for each component weight, considering the conditional probabilities of a match (m) and nonmatch (u). The probabilities of the random vector $\gamma = (\gamma_1, ..., \gamma_n)$ conditional on the match status Z are defined as

$$u_\gamma = P(\gamma|Z = 0)$$
$$m_\gamma = P(\gamma|Z = 1)$$

where $Z = 0$ stands for nonmatch and $Z = 1$ stands for match. Specifically, the vector is formed by summing up the weights of all components, represented as:

$$w(\gamma^k) = w^1 + w^2 + \cdots + w^k$$

where w^k is a vector of weights, and w^1 represents the weight of component one (e.g., last name). Each component weight w^k is calculated as a log odds ratio:

$$w^k(\gamma^k) = \log\left(\frac{P(\gamma|Z = 1)}{P(\gamma|Z = 0)}\right) = \log m(\gamma^k) - \log u(\gamma^k)$$

Here, m stands for the likelihood of a match, u represents the likelihood of a nonmatch, and γ^k denotes the value of the comparison components of variable k. These values are specific to each entry in the record and are influenced by the frequency of the entry and the degree of error in the field. Thus, these weights play a crucial role in distinguishing between matches and nonmatches in the context of record linkage.

When assessing the likelihood of a match between records, record linkage adjusts for common or certain rare entries in the dataset since some entries, e.g., last names, vary in frequency. For instance, a common last name like “Smith” is likely to be found in both records, so a match would have a lower weight compared to a less common surname like “Raab” (Osborne & McLaurin, 2006).

The formulas for calculating the relevant values of m and u depend on the match status such that for a match it is

$$m = \frac{f_j}{N}(1 - 2e_T - 2e_0)$$

$$u = \left(\frac{f_j}{N}\right)^2 (1 - 2e - e_T - 2e_0)$$

And for a nonmatch it is

$$m = 2e + e_T$$

$$u = [1 - (2e - e_T)\sum_j \left(\frac{f_j}{N}\right)^2](2e_0)$$

Although we didn’t rely on names, when a last name was missing on either record, the record linkage algorithm is designed to apply

$$m = 2e_0$$
$$u = 2e_0$$

f_j in all previous equations is the frequency of a particular entry, N is the total number of entries in the dataset. e is the probability of a name being misreported, and e_0 is the probability that a record is missing. These probabilities are calculated separately for smaller populations. Additionally, e_T is the probability of a name being reported differently in either dataset. However, due to challenges in distinguishing this term from e, it isn’t calculated separately.

Finally, note that the set of algorithms to be used in the counterfactual identification come from the R Project’s RecordLinkage package (2023), which were originally designed to detect errors among databases. The R Project algorithms offer a robust approach to identifying the “same” student across spreadsheets, but with inconsistencies in, for example, their names or dates of birth. Due to space limitations, see Sariyar and Borg (2010) for details on the packaged algorithms to be used. In this study, we extended the use of these machine learning and classification algorithms to the counterfactual and potential outcomes frameworks.

As Figure 1 indicates, instead of identifying errors in databases as originally designed (Dusetzina et al., 2014), the framework shown in that figure essentially: searched for participants who look the closest across two datasets (Panels (a) and (b) in Figure 1), evaluated all feasible

set of matches across the total number of indicators via RecordLinkage (Panel (c)), and extracted the most optimal solutions—those above the 0.75 threshold, as shown in Panel (d), also in Figure 1. Recall that the higher these feasibility weights are, the better the counterfactual identification is (see Dusetzina et al., 2014; Sariyar & Borg, 2010).

Data sources

The National Education Longitudinal Study (NELS, 1988–2000) and the Educational Longitudinal Study (ELS, 2002–2012) configure the main source of data. Both samples are representative of the population of U.S. high school students transitioning to college and were measured exactly one decade apart, which enables us to both have a secondary and independent source to assess for robustness and sensitivity of our analyses and offers the possibility of reaching more nuanced understandings of these students. Although these datasets offer thousands of indicators (see this Table for an example of some indicators used in a recent study by González Canché, 2020), for our study we first selected academic indicators related to math and science (see Tables 4 and 5, Panels B), given their relevance for success in STEM (National Research Council, 2011), and then added individual attributes including gender, ethnicity, and an index of socioeconomic status (see Tables 4 and 5, Panels A). In both the NELS and ELS samples, this index was created as a function of parents' (or guardians') employment prestige, level of education, and salary. Also, in both samples these SES indexes were normalized to range to about -2 to 2, with negative values signaling socioeconomic hardship and positive values indicating socioeconomic prosperity.

Variable selection for probabilistic matching

Initially, we aimed to include both high school academic indicators and individual attributes as part of the counterfactual identification process. This initial decision was based on our goal of creating as many comparable matches as possible; however, we were failing to see what all previous literature on community college and four-year students has pointed out, that these students are systematically different in both their academic indicators and their attributes (Doyle, 2009; González Canché, 2020; Melguizo, et al., 2009; Reynolds & DesJardins, 2009). Accordingly, trying to match in all of them would result in identifying community college students who mirror their four-year counterparts in academic and sociodemographic attributes. Consequently, when we applied these matching techniques using both academic and sociodemographic indicators, the probability weights identified fewer than 90 community college cases in each sample. The reason for this lack of matches was that four-year students graduating from STEM tend to be White, Asian, male, and wealthier, compared to community college entrants. In other words, "forcing" the matching process to include sociodemographic indicators resulted in basically lack of matches for few students in the community college truly match four-year students socioeconomic and demographic indicators. This is why we decided to limit the matched to only their high school academic indicators, including cumulative grade point average (GPA), number of credits in math and science, and number of math and science courses their high schools require for them to graduate (Tables 4 and 5, Panels B for the NELS and ELS samples, respectively). This strategy resulted in more matches, who were counterfactually the same academically speaking but didn't look the same in gender, race and socioeconomic standing. The results of these procedures can be seen in Figures 2 and 3, and Table 1, as discussed in our findings section.

Post-identification analyses

All two-year entrants who meet the feasibility weight threshold were assigned a value of 1 (i.e., counterfactually matched), and their nonmatched counterparts were assigned a value of 0 (i.e., nonmatched). Each two-year participant, regardless of matched status, also has an observed value (see Panel (d) in Figure 1). With this information we estimated the expected probability of a four-year degree attainment in STEM across matched and unmatched participants. These results can be observed in Table 2. Moreover, since community college students can, by design, attain less than four-year degrees, we also estimated their probabilities of attaining a less-than-four-year STEM degree, as can also be seen in Table 2.

Table 3 shows an index we're calling *realization rate.* We created this realization rate index from two indicators: each student's observed STEM outcome and feasibility weight. These two values were multiplied so that if a student attained a BS in STEM (i.e., an outcome with a value of 1) and had feasibility weight of 1, then this student would have a value of 1 in this index. If a student had a feasibility of 0.75 and obtained a BS in STEM, this student would retain this feasibility weight of 0.75 in this index. This implies that, to the extent matched students are obtaining BS degrees in STEM, on average we should expect them all to have realization rate indexes over 0.75.

Note that we also computed this index for nonmatched students, who may or may not have attained a BS degree in STEM. In this case, we changed their matched indicator to 1 in order to avoid multiplying their observed outcomes by zero and thus not being able to compute their *realization rate* indexes. In this case, since nonmatched students had feasibility weights below 0.75, we observed realization

indexes lower than their matched counterparts. As shown in Table 3, to further compare these performances, we also took the difference of the average realization rate indexes by matched and nonmatched participants from their feasibility indexes. We did this to assess the degree loss rate. That is, imagine that among matched participants the realization rate is 0.30 on average, but their average feasibility weight is 0.80. In this case, the gap between the feasibility weight (i.e., their expected probability of counterfactually having attained a four-year degree in STEM) and their cohort average BS attainment rate is 0.50 (or 0.80 – 0.30). This value may then be read as having decreased their mean probabilities of attaining a BS degree in STEM by 50 percentage points. In other words, we lost the possibility of increasing the production of STEM degrees among two-year entrants by 50 percentage points, as we further elaborate in the findings section. A similar set of analyses will be computed for nonmatched participants.

Finally, we implemented generalized boosted models (GBM) via Gradient Boosting Machine (see Ridgeway, 2024) to identify drivers of BS in STEM degree attainment among counterfactual community college students. Essentially, via boosting, GMB aims to minimize some loss or misclassification via negative gradient (or gradient descent) when the classification is less than optimal or positive gradient (or gradient ascent) if the classification is above the best loss fit (i.e., perfect classification).

An important conceptual characteristic of GBM with statistical implications is that GBM offers the best prediction (or prediction that minimizes loss or inaccuracy in predictions) based on the values of X (i.e., the variables and attributes of the units of analysis) for which we have observations (Ridgeway, 2024). This implies two things. One, we don't know if our model is the best model because there may be many other important Xs that we don't have access to in our datasets—hence preventing us from claiming causality. In this study these Xs are academic and nonacademic (i.e., sociodemographic indicators) attributes observed. The second implication is that we can explore how the combination of these indicators behaves with multiple GMB model specifications—i.e., adding or removing some Xs from the model. Specifically, in the full models shown in Figures 4 and 5, we included all academic and nonacademic indicators, whereas in the subsequent two models in each figure we excluded SES and retained only gender and ethnicity as indicators of interest, respectively.

A concern related to the use of boosting is the possibility of overfitting, but in GBM this concern is minimized by imposing λ, a learning rate through shrinkage (Friedman, 2001). Boosting learns from weak models until it reaches the optimal minimum loss rate through T number of iterations (Ridgeway, 2024) via cross-validation by selecting random subsamples without replacement of the training datasets. Note that training datasets consist of randomly taking 70% of the analytic samples in both NELS and ELS.

Once the cross-validation has gone over all T iterations, we are then able to identify the most important drivers for regression or classification. In our case, since the outcomes are four-year degree attainment in STEM, we relied on GBM models for classification. In GBM the influence of a variable X_j is measured by the empirical improvement (loss reduction) associated with splitting trees on specific values of X_j during the training iterations T (Ridgeway, 2024). The results of Xs relevance is shown in Figures 4 and 5, wherein in our findings section we will further explore the meaning of these indicators with respect to their impact on the outcome of interest—i.e., four-year STEM degree attainment.

Finally, an important attribute of GBM is its capability to account for survey weights during the classification or regression processes. This is important for our efforts to offer estimates that are nationally representative of the United States across two different decades.

Findings

For clarity, we have organized this section by research question.

Can we develop an innovative method using counterfactual causality and predictive analytics to identify two-year entrants highly likely to attain a four-year STEM degree?

Our identification strategy successfully identified community college students with a high probability of attaining a STEM degree across different analytic samples. Let us note that these probabilities were obtained from counterfactually matching community college students with individuals who pursued and obtained a bachelor's degree in STEM fields after starting in four-year institutions.

Figures 2 and 3 show the counterfactual probabilities distributions of two-year entrants in the NELS and ELS samples. Our identification threshold was 0.75, highlighted by the vertical lines. Students positioned to the right of these vertical lines were classified as counterfactual cases. Notably, students to the left of this threshold serve as comparison units. Additionally, we observed a concentration of cases around the 0.675 mark, indicating a density and frequency bump near our threshold in both samples. This observation raises noteworthy issues, because students closer to the threshold are likely to exhibit more similar outcomes than those further away. Consequently, our results may be conservative. Future studies could explore alternative techniques, such as quantile regressions, to better understand the impact of deviations from our selected threshold on the identification of counterfactual probabilities.

After operationalization, what is the prevalence of these counterfactual cases in the two latest nationally representative samples of U.S. students transitioning from high school to college?

Table 1 provides additional context to Figures 2 and 3. This table also answers the second question of this study, aiding in the examination of the prevalence of these counterfactual cases across the two analytic samples. This table contains two panels, each corresponding to one of the analytic samples. Within each panel, we present the distribution of students identified as counterfactuals alongside the distribution of their nonmatched community college peers.

A comparative analysis spanning across decades reveals an appreciable increase in the number of students in the ELS sample, the most recent cohort, by approximately 100,000. Notably, the matched sample of this latter cohort featured over 100,000 more cases. Thus, both samples exhibited fewer than 777,000 nonmatched cases. However, in the NELS sample, about 102,000 students were matched, whereas this number increased to 229,000 cases in the ELS sample. In terms of percentages, the NELS sample indicated that about 12% of students were identified as counterfactuals, while this proportion surged to 23.7% in the ELS sample.

These results suggest that our identification strategy consistently pinpointed more than 100,000 students with a high likelihood of attaining a bachelor's degree in STEM, with this representation doubling in the following dataset.

What is the distribution of observed four-year STEM degree attainment among identified counterfactual and non-counterfactual two-year entrants? Also, how do distributions for less than four-year STEM degrees differ in these groups, considering variations across decades?

Table 2 present a descriptive correlational analysis of how our counterfactual identification strategy correlates with the probabilities of attaining a four-year degree in STEM. Each table consists of a two-by-two cross tabulation, delineating two groups and two outcomes.

In the first group, we examine non-counterfactual cases and their outcomes regarding four-year STEM degree attainment. Conversely, in the second group, we present these outcomes distributed among matched students. The associated probability value reflects a chi-squared distribution test that assesses the presence of an association between the two variables.

Across both samples, consistent results emerge regarding four-year STEM degree attainment among nonmatched students. Nonmatched community college students exhibit a likelihood of less than 6.4% in attaining a four-year degree, with a downward trend evident in the most recent sample, where the probability drops to 5.4%. On the contrary, matched students demonstrated substantially higher probabilities of obtaining a four-year degree than their unmatched counterparts. In the NELS sample, these probabilities reached 18.4%, albeit decreasing to 14% in the more recent ELS sample. This indicates that, using the same analytic techniques and identification procedures, counterfactual community college students are experiencing lower success rates than their matched peers from a decade earlier.

What is the distribution of these estimates when considering less-than-four-year degrees in STEM by these comparison groups?

The rows labeled "STEM degree, Less than BS" in Table 2 address this question. It is evident that, across decades, about 14% of community college students who weren't matched attained a degree in STEM. In the case of matched students who didn't attain a BS degree in STEM, there is a notable three-percentage-point decrease in the attainment of less than a BS degree in STEM in the most recent sample. Specifically, in the NELS sample, 20% of counterfactually matched community college students attained a less-than-four-year STEM degree. In the ELS sample this percentage was 17.3 percent.

Although our analysis doesn't directly measure the impact of our identification strategy on increasing less-than-four-year BS degrees in STEM, our matched students tended to outperform their nonmatched peers in this outcome. However, this advantage shrank across decades, indicating potential shifts or changes in educational pathways and attainment patterns.

How do the observed outcomes of community college students differ from their expected probabilities in achieving both four-year and less-than-four-year STEM degrees? Additionally, what are the variations in success rates between counterfactual and non-counterfactual community college entrants?

To address these questions, we present the realization rates discussed in the methods section. The results of these procedures are shown in Table 3. Panel A shows the realization rate of NELS students by matched status. Overall, we see a significantly higher realization rate for matched students, averaging at 0.153 (standard deviation = 0.32), compared with a mere 0.035 (standard deviation = 0.14) among nonmatched students. Additionally, Table 3 shows the probability or feasibility weight, with mean values of 0.836 (standard deviation = 0.08) for matched students and 0.501 (standard deviation = 0.20) for nonmatched students. These latter values represent the "ideal" scenario in which all students attained a four-year degree in STEM.

From this perspective, then, the gap between matched students' feasibility index and realization rate may be read as a form of STEM loss rate in this sample. Specifically, by subtracting the average feasibility weight from the realization

rate (0.8358 – 0.1526 = 0.6832), we estimate a gap of 68 percentage points in STEM attainment probability. This signifies that, without intervention, matched students in the NELS sample experienced a substantial decrease in the probability of attaining a BS in STEM. The loss gap for nonmatched students was comparatively lower, measuring at 0.466 (or 0.5013–0.0353). Although this latter gap was expected due to the inherently lower probabilities of nonmatched students, as indicated by their feasibility indexes (up to 0.749), their attainment rates were also lower, as seen in Table 2.

Similar patterns persist for nonmatched students in the ELS sample, a decade later. Here, the mean realization rate was 0.037 (standard deviation = 0.14) compared to 0.0353 in the NELS sample. For matched participants, the mean realization rate was lower, decreasing from 0.1526 (standard deviation = 0.32) to 0.1169 (standard deviation = 0.29). These worsening cases across all analyses corroborates the widening loss rate of BS STEM completion among community college students identified as four-year counterfactuals in recent samples. In this case, the gap for these counterfactually matched students widened to 0.720 (or 0.8366–0.1169).

How do the previous estimates change when measuring the attainment of a less-than-four-year degree in STEM?

In the case of less than a four-year BS in STEM degree, we continue to see a deterioration in outcomes in the most recent sample. Specifically, in the NELS case, counterfactual community college students had a realization rate of 0.172 (standard deviation = 0.36), whereas a decade later, this rate decreased to 0.125 (standard deviation = 0.30). Contrarily, the realization rate of nonmatched students remained consistent across decades, hovering around a mean value of about 0.08 (standard deviation ≈ 0.20).

What are the best predictors of success for counterfactual two-year students while considering BS and less than BS degrees in STEM?

To begin addressing this question, we'll delve into Tables 4 and 5, which delineate the levels of each indicator disaggregated by counterfactually matched status.

In the NELS sample (see Table 4), we see a relatively balanced gender representation, with women constituting about 50% of nonmatched students. However, among matched community college entrants, women were slightly overrepresented, comprising about six percentage points more than their male counterparts. Regarding ethnicity, white students dominate both nonmatched and matched groups, comprising about 70% and 82%, respectively. Asian representation remains low in both categories, not surpassing 4.3%—this may indicate a tendency for Asian students in the NELS sample to bypass starting college in the public two-year sector. Similarly, Black students weren't prevalent in the matched sample, representing only 3.7%. Hispanic students, however, constitute the second-highest represented group among matched students (8.6%). Hispanic students were also the second highest represented group among nonmatched students with 14%, potentially indicative of their propensity to start college in the community college sector.

In Table 5, which contains the attributes of ELS students, women became even more represented in the matched sample, comprising about 60%. The representation of white students decreased in the matched sample from 82% to 71.3% over the decade. Asian students remain a small portion of the total (3.6%). Notably, both black and Hispanic student representation almost doubles in the matched sample, reaching 7.6% and 15.6%, respectively. Although the ELS sample provides more detailed accounts of ethnicity, including multi-race and Native and Alaska Native, their representation in the sample remained low. With this contextualization of our participants' attributes in mind, we proceed to identify the most important predictors of BS degree attainment in STEM.

To formally address these questions, we employ six GBM specifications. These models enable the identification of the most important and stable drivers for STEM four-year degree completion. As outlined in the methods section, we first trained the data using 70% of each unit in each sample and then tested the GBM performance with the remaining 30% in each sample. The results are shown in Figures 4 and 5. Each figure contains three GBM specifications. The first specification, termed the full model, contains all attributes discussed in Tables 4 and 5, in addition to academic achievement used for generating probabilistic weights or feasibility index as well as the socioeconomic (SES) index provided by NELS and ELS. The second specification excluded SES from these models, The third specification only includes the attributes contained in Tables 4 and 5.

Figure 4 corresponds to the NELS sample. In the full GBM, socioeconomic status (SES) emerges as the most influential factor for BS attainment in STEM, followed by GPA. Subsequently, the number of courses required in science during high school, along with the number of units in math and science attained during high school are identified as the third most important drivers. Among nonacademic attributes, Asian ethnicity is the first to surface, followed by gender (woman), and Hispanic ethnicity. When SES is omitted from the model, GPA takes precedence as the most critical predictor, followed by the number of high school math courses, then by the number of science requirements, and finally by the number of science units completed. Among nonacademic indicators, gender (woman) surpasses Asian ethnicity in importance.

In the final GBM, which excluded all academic indicators, Hispanic ethnicity was the most important driver, followed by gender (woman). It's worth noting that Hispanic students exhibited about 15.15 percentage points higher probabilities of attaining a STEM BS degree compared to non-Hispanics. Additionally, women demonstrate probabilities of attaining a BS in STEM approximately 6.3% higher than men.

The ELS sample (see Figure 5) presented some similarities and differences. When socioeconomic status was included in the full model, it remained the most important driver, followed by GPA. These findings mirrored the NELS GBM results. Similarly, when socioeconomic status was excluded, GPA became the most predominant driver.

In the ELS sample, the single most consistent nonacademically related indicator was gender (woman), and when only students' demographics were considered in the GBMs, women and Hispanic ethnicity were the top two drivers. Notably, despite the relevance of these indicators, our correlational analyses indicated that Hispanic students were actually five percentage points less likely to attain a BS in STEM than non-Hispanic students. Similarly, women were 1.3 percentage points less likely to attain this degree than men. These results underscore the limited precision and strength of association between these demographic indicators and the outcome, despite their apparent importance in reducing prediction error during the training process.

In summary, these results indicate that in the most recent sample, demographic attributes have become less important as drivers of academic success compared to the previous sample.

Limitations

Despite the insights provided by this study, several limitations should be acknowledged. First, the study focused primarily on demographic and academic predictors of success, overlooking other potential factors, such as social support networks, motivation and career aspirations. Additionally, to our knowledge, this is the first study that employs probabilistic matching to identify counterfactual cases. Research on how or the degree to which omitted variable bias may impact our results remains to be determined. Moreover, as stated above, our cut off point of 75% may be expanded to use other regression methods like regression discontinuity or quantile regression to offer alternative modeling approaches to estimate the impact of our counterfactual identification. Finally, qualitative research methods could provide deeper insights into the experiences and perspectives of counterfactually identified community college students in STEM education.

Discussion

Community colleges play a crucial role in higher education, serving as accessible pathways for many students seeking to pursue degrees in STEM fields. However, achieving success in STEM programs for students starting at community colleges remains a complex and multifaceted endeavor, influenced by factors ranging from academic preparation to demographic characteristics. Our research findings should serve as a basis for data-driven decision-making, prompting tailored support systems that address the unique needs of community college students aspiring to transfer to four-year STEM programs.

Our analyses highlight the influence of both demographic and academic factors on the success of community college students in STEM education. Gender and ethnicity emerged as significant predictors, with female students consistently driving success across various samples, highlighting the need to address gender disparities in STEM fields. Similarly, while Hispanic ethnicity showed higher probabilities of attaining a STEM BS degree in some analyses, correlational findings revealed lower actual attainment rates. This indicates the complexity of demographic influences on academic outcomes.

Socioeconomic status emerged as a critical predictor in the full GBM specification, underscoring the impact of socioeconomic factors on educational outcomes. However, when socioeconomic was excluded from the model, GPA emerged as the most influential predictor, suggesting that academic achievement mitigated the effects of socioeconomic disparities—which corroborates the relevance of providing low income students with quality education opportunities. Academic achievement, as measured by GPA and high school coursework, also played a pivotal role in predicting success in STEM education. Our analysis demonstrated that GPA was consistently identified as a significant predictor across different models, emphasizing the importance of academic performance in determining outcomes. Additionally, the number of science courses required during high school and the number of units in math and science completed were also significant factors influencing success in STEM programs. These findings underscore the importance of supporting students from diverse socioeconomic backgrounds and providing resources to ensure equitable access to educational opportunities.

The findings of this study have important implications for practice and policy in STEM education. Although our identification framework effectively identifies cases with elevated probabilities of attaining BS degrees in STEM, we have noted a troubling decline in the likelihood of these students obtaining their four-year degrees in STEM. The disparity between the anticipated likelihood of attaining

a STEM degree and the actual outcomes underscore a critical issue in the American higher education system. Community college students harboring the potential to excel in STEM fields may encounter formidable barriers due to the absence of adequate resources, support and policies. This not only results in a palpable loss of human capital but also diminishes the diversity within the STEM workforce. The repercussions extend to the career and financial well-being of these students, as delays in education completion can hinder their professional trajectory and earning potential. Based on our results, interventions and programs that support community colleges and their STEM programs may reduce educational disparities.

Conclusions

The focus of our study on four-year STEM degrees marks an important departure from existing literature on the baccalaureate gap and transfer effects. Our goal is to provide insights into opportunities that reduce social, academic, and economic disparities by identifying, as early as possible, minoritized two-year students with high probabilities of a STEM four-year degree attainment.

In conclusion, our study offers a counterfactual framework for affordable human-resource-formation in STEM, using the community college sector as an effective entry point. While our identification framework successfully identifies cases with higher probabilities of attaining BS degrees in STEM, we have observed a concerning decrease in the likelihood of these students attaining their four-year degrees in STEM. This reduction highlights the urgency to design efficient and effective plans of action to reduce failure rates. Although we show that our identification was successful in identifying cases with higher probabilities to attain BS in STEM, we also found a decrease of almost 70 percentage points in these students' prospects of attaining four-year degrees in STEM. This loss potentially represents a foregone opportunity to improve the mobility prospects of these students based on the additional wage premium that would be earned with further education.

As highlighted in our literature review, transitioning from a community college to a four-year degree not only offers a more affordable route compared to the traditional four-year path (González Canché, 2014, 2022) but also may hold both greater significance for upward mobility, particularly for the typically minoritized and economically disadvantaged students who are more prevalent in this sector than the four-year college sector overall. The results in this point suggest education disparities may be reduced if the community college system is better leveraged as a source of four-year STEM degrees.

References

Clark, B. R. (1960). *The open door college: A case study*. McGraw-Hill.

Cohen, A. M. (1987). *Facilitating degree achievement by minorities: The community college environment.*

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

Dougherty, K. J. (1992). Community colleges and baccalaureate attainment. *The Journal of Higher Education*, *63*, 188–214. doi:10.2307/1982159.

Doyle, W. R. (2009). The effect of community college enrollment on bachelor's degree completion. *Economics of Education Review*, *28*(2), 199–206.

Dusetzina S., Tyree S., & Meyer, A. M. (2014). *Linking data for health services research: A framework and instructional guide.* Agency for Healthcare Research and Quality (US). Available from: https://www.ncbi.nlm.nih.gov/books/NBK253312/.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5):1189–1232.

Garrison, R. H. (1975). Community colleges: The literature of the two-year college. *Change: The Magazine of Higher Learning*, *7*(3), 58–60.

González Canché, M. S. (2014). Is the community college a less expensive path toward a bachelor's degree? Public 2- and 4-year colleges' impact on loan debt. *The Journal of Higher Education*, *85*, 723–759. doi:10.1353/jhe.2014.0026.

González Canché, M. S. (2017a). Financial benefits of rapid student loan repayment: An analytic framework employing two decades of data. *The ANNALS of the American Academy of Political and Social Science, 671*(1), 154–182. https://journals.sagepub.com/doi/full/10.1177/0002716217701700.

González Canché, M. S. (2017b). Community college scientists and salary gap: Navigating socioeconomic and academic stratification in the U.S. higher education system. *Journal of Higher Education, 88*(1), 1–32. https://www.tandfonline.com/doi/full/10.1080/00221546.2016.1243933.

González Canché, M. S. (2018). Reassessing the two-year sector's role in the amelioration of a persistent socioeconomic gap: A proposed analytical framework for the study of community college effects in the big and geocoded data and quasi-experimental era. In M. B. Paulsen (Ed.), *Higher education: Handbook of theory and research.* Springer. https://www.springer.com/gp/book/9783319268286.

González Canché, M. S. (2020). Community college students who attained a 4-year degree accrued lower student loan debt than 4-year entrants over 2 decades: Is a 10 percent debt accumulation reduction worth the added "risk"? If so, for whom? *Research in Higher Education*. https://doi.org/10.1007/s11162-019-09565-9.

González Canché, M. S. (2022). Post-purchase Federal Financial Aid: How (in) Effective is the IRS's Student Loan Interest Deduction (SLID) in Reaching Lower-Income Taxpayers and Students? *Research in Higher Education*, *63*(6), 933.

González Canché, M. S., Lee, J. C., Harding, J. L., Turk, J. M., Bae, J. Y., & Zhang, C. (2023). Post-Baccalaureate Federal Loans De-Subsidization: Impacts on Compositional Attributes, Extensive and Intensive Borrowing Margins, and Anticipatory Effects. *The Journal of Higher Education*, 1–38.

Helland, P. C. (1987). *Establishment of public junior and community colleges in Minnesota, 1914-1983*. Minnesota Community College System.

Hu, X., Ortagus, J. C., & Kramer, D. A. (2017). The community college pathway: An analysis of the costs associated with enrolling initially at a community college before transferring to a 4-year institution. *Higher Education Policy*, 1–22.

Lewis, D. (1974). Causation. *The Journal of Philosophy*, *70*(17), 556–567.

Lichtenberger, E., & Dietrich, C. (2017). The community college penalty? examining the Bachelor's completion rates of community college transfer students as a function of time. *Community College Review*, *45*(1), 3–32.

Malcom, S. & Feder, M. (2016). Barriers and opportunities for 2-year and 4-year STEM degrees: Systemic change to support students' diverse pathways. *National Academies of Sciences, Engineering, and Medicine.* https://www.nap.edu/read/21739.

Melguizo, T., Kienzl, G. S., & Alfonso, M. (2011). Comparing the educational attainment of community college transfer students and four-year college rising juniors using propensity score matching methods. *The Journal of Higher Education*, *82*(3), 265–291.

Monaghan, D. B., & Attewell, P. (2015). The community college route to the bachelor's degree. *Educational Evaluation and Policy Analysis*, *37*(1), 70–91.

National Research Council. (2011). *Successful K-12 STEM education: Identifying effective approaches in science, technology, engineering, and mathematics*. National Academies Press.

National Student Clearinghouse (February 2022). National Six-Year Completion Rate Reaches 62.2%. National Student Clearinghouse. Available from https://www.studentclearinghouse.org/national-six-year-completion-rate-reaches-62-2/.

National Science Board. (2021). The STEM Labor Force of Today: Scientists, Engineers, and Skilled Technical Workers. National Science Board. Available from https://ncses.nsf.gov/pubs/nsb20212/figure/LBR-12.

National Science Foundation. (2021). Survey of Doctorate Recipients (SDR). Available from https://ncses.nsf.gov/surveys/doctorate-recipients/2021.

Ridgeway, G. (2024). Generalized Boosted Models: A guide to the GBM package. *Update* 2024. Available from https://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*(469), 322–331.

Sariyar, M., & Borg, A. (2010). The RecordLinkage package: Detecting errors in data. *The R Journal*, *2*(2), 61–67.

Shapiro, D., Dundar, A., Huie, F., Wakhungu, P. K., Yuan, X., Nathan, A., & Hwang, Y. (2020). Tracking transfer: Measures of effectiveness in helping community college students to complete bachelor's degrees.(signature report no. 13, updated 2020). *National Student Clearinghouse*. https://nscresearchcenter.org/wp-content/uploads/Sig13Update2020_Fall2013Cohort.xlsx.

Spindt, H. A. (1957). Beginnings of the junior-college in California, 1907-1921. *College and University*, *33*(1), 22–28.

Vaughan, G. B. (1995). *The community college story: A tale of American innovation*. American Association of Community Colleges, National Center for Higher Education.

Tables and figures

TABLE 1. HOW PREVALENT ARE THEY?

Matching Status	Levels	n	%
NELS Nationally Representative Sample			
Matched	0	768888	88.3
	1	102306	11.7
	all	871194	100.0
ELS Nationally Representative Sample			
Matched	0	740066	76.3
	1	229908	23.7
	all	969974	100.0

Note: Unmatched is no counterfactual match (<.75). Matched is a counterfactual match (≥ .75).

TABLE 2. NATIONALLY REPRESENTATIVE SAMPLES: OUTCOME INDICATORS

Variable	Levels	Unmatched	$\%_0$	Matched	$\%_1$	n_{all}	$\%_{all}$
NELS Sample (1988 – 2000)							
BS STEM degree	No	720623	93.7	83655	81.8	804278	92.3
	Yes	48265	6.3	18651	18.2	66916	7.7
$p < 0.0001$†	all	768888	100.0	102306	100.0	871194	100.0
STEM degree, Less than BS	No	577044	85.8	65092	80.0	642136	85.2
	Yes	95319	14.2	16291	20.0	111610	14.8
$p < 0.0001$†	all	672363	100.0	81383	100.0	753746	100.0
ELS Sample (2002 – 2012)							
BS STEM degree	No	699924	94.6	197660	86.0	897584	92.5
	Yes	40142	5.4	32248	14.0	72390	7.5
$p < 0.0001$†	all	740066	100.0	229908	100.0	969974	100.0
STEM degree, Less than BS	No	601844	86.0	163528	82.7	765372	85.3
	Yes	98080	14.0	34132	17.3	132212	14.7
$p < 0.0001$†	all	699924	100.0	197660	100.0	897584	100.0

†Chi-Squared test

TABLE 3. OUTCOMES AND PREDICTED OUTCOMES QUESTION 4

Variable	Matched	n	$\bar{x}$	St.Dev.	Min	Max
Panel A: NELS Sample						
BS Realization rate	No	766302	0.0353	0.1429	0.0000	0.7494
	Yes	101665	0.1526	0.3235	0.0000	1.0000
$p < 0.0001$	all	867967	0.0491	0.1781	0.0000	1.0000
Lower than BS Realization rate	No	672363	0.0772	0.2026	0.0000	0.7494
	Yes	81383	0.1718	0.3455	0.0000	1.0000
$p < 0.0001$	all	753746	0.0874	0.2244	0.0000	1.0000
Weight	No	768888	0.5013	0.1953	0.0654	0.7494
	Yes	102306	0.8358	0.0754	0.7645	1.0000
$p < 0.0001$	all	871194	0.5406	0.2143	0.0654	1.0000
Panel B: ELS Sample						
BS Realization rate	No	740066	0.0337	0.1428	0.0000	0.7343
	Yes	229908	0.1169	0.2909	0.0000	1.0000
p < 0.0001	all	969974	0.0535	0.1920	0.0000	1.0000
Lower than BS Realization rate	No	740066	0.0821	0.2134	0.0000	0.7343
	Yes	229908	0.1251	0.3014	0.0000	1.0000
$p < 0.0001$	all	969974	0.0923	0.2379	0.0000	1.0000
Weight	No	740066	0.6100	0.1094	0.0700	0.7343
	Yes	229908	0.8366	0.0807	0.7610	1.0000
$p < 0.0001$	all	969974	0.6637	0.1412	0.0700	1.0000

TABLE 4. NELS NATIONALLY REPRESENTATIVE SAMPLE: DEMOGRAPHIC CHARACTERISTICS

Variable	Levels	n_0	$\%_0$	n_1	$\%_1$	n_{all}	$\%_{all}$
Panel A. Sociodemographic Attributes							
woman	0	389422	50.6	47986	46.9	437408	50.2
	1	379466	49.4	54320	53.1	433786	49.8
$p < 0.0001$	all	768888	100.0	102306	100.0	871194	100.0
white	0	233192	30.4	18768	18.3	251960	29.0
	1	534522	69.6	83538	81.7	618060	71.0
$p < 0.0001$	all	767714	100.0	102306	100.0	870020	100.0
asian	0	740944	96.5	97861	95.7	838805	96.4
	1	26770	3.5	4445	4.3	31215	3.6
$p < 0.0001$	all	767714	100.0	102306	100.0	870020	100.0
black	0	680948	88.7	98509	96.3	779457	89.6
	1	86766	11.3	3797	3.7	90563	10.4
$p < 0.0001$	all	767714	100.0	102306	100.0	870020	100.0
hispanic	0	660545	86.0	93511	91.4	754056	86.7
	1	107169	14.0	8795	8.6	115964	13.3
$p < 0.0001$	all	767714	100.0	102306	100.0	870020	100.0

Variable	Matched	n	$\bar{x}$	St.Dev.	Min	Max
Panel B. Academic Indicators						
units_in_math	0	620127	2.6694	1.1103	0.0000	6.0000
	1	102306	3.4129	0.8021	0.0000	6.0000
$p < 0.0001$	all	722433	2.7747	1.1030	0.0000	6.0000
units_in_science	0	620127	2.4134	0.9872	0.0000	8.0000
	1	102306	3.0090	0.7471	1.0000	5.0000
$p < 0.0001$	all	722433	2.4977	0.9791	0.0000	8.0000
gpa_all _courses[†]	0	486255	13.4171	27.2215	0.0000	103.0800
	1	102306	3.4187	5.4261	1.8300	86.0000
$p < 0.0001$	all	588561	11.6791	25.1332	0.0000	103.0800
math_important	0	768888	0.8350	0.3712	0.0000	1.0000
	1	102306	0.8933	0.3087	0.0000	1.0000
$p < 0.0001$	all	871194	0.8418	0.3649	0.0000	1.0000
absorbed_by_math	0	768888	0.5960	0.4907	0.0000	1.0000
	1	102306	0.6391	0.4803	0.0000	1.0000
$p < 0.0001$	all	871194	0.6011	0.4897	0.0000	1.0000
hs_math_reqmnt	0	542393	2.4251	0.6079	0.0000	4.0000
	1	93487	2.3127	0.5708	0.0000	4.0000
$p < 0.0001$	all	635880	2.4086	0.6039	0.0000	4.0000
hs_science_ reqmnt	0	542173	2.1180	0.5640	0.0000	4.0000
	1	93358	2.1397	0.5614	0.0000	4.0000
$p < 0.0001$	all	635531	2.1212	0.5637	0.0000	4.0000

Note: n_0 reflects no counterfactual match, n_1 is a counterfactual match.
[†]Cumulative GPA may exceed 100 percent in NELS because of quality of courses.

TABLE 5. ELS NATIONALLY REPRESENTATIVE SAMPLE: PARTICIPANTS' DEMOGRAPHIC CHARACTERISTICS

Variable	Levels	n_0	$\%_0$	n_1	$\%_1$	n_{all}	$\%_{all}$
Panel A. Sociodemographic Attributes							
woman	0	361255	48.8	92545	40.2	453800	46.8
	1	378811	51.2	137363	59.8	516174	53.2
$p < 0.0001$	all	740066	100.0	229908	100.0	969974	100.0
white	0	344008	46.5	66068	28.7	410076	42.3
	1	396058	53.5	163840	71.3	559898	57.7
$p < 0.0001$	all	740066	100.0	229908	100.0	969974	100.0
asian	0	711033	96.1	223884	97.4	934917	96.4
	1	29033	3.9	6024	2.6	35057	3.6
$p < 0.0001$	all	740066	100.0	229908	100.0	969974	100.0
black	0	626571	84.7	212493	92.4	839064	86.5
	1	113495	15.3	17415	7.6	130910	13.5
$p < 0.0001$	all	740066	100.0	229908	100.0	969974	100.0
hispanic	0	576216	77.9	194055	84.4	770271	79.4
	1	163850	22.1	35853	15.6	199703	20.6
$p < 0.0001$	all	740066	100.0	229908	100.0	969974	100.0
multi_race	0	708401	95.7	224918	97.8	933319	96.2
	1	31665	4.3	4990	2.2	36655	3.8
$p < 0.0001$	all	740066	100.0	229908	100.0	969974	100.0
nat_am_alaska	0	734101	99.2	228122	99.2	962223	99.2
	1	5965	0.8	1786	0.8	7751	0.8
$p = 0.17$	all	740066	100.0	229908	100.0	969974	100.0

TABLE 5. ELS NATIONALLY REPRESENTATIVE SAMPLE: PARTICIPANTS' DEMOGRAPHIC CHARACTERISTICS (CONTINUED)

Variable	Matched	n	$\bar{x}$	St.Dev.	Min	Max
Panel B. Academic Indicators						
units_in_math	0	740066	2.8404	1.1398	0.0000	7.0000
	1	229908	3.5414	0.7463	0.0000	6.0000
$p < 0.0001$	all	969974	3.0065	1.1009	0.0000	7.0000
units_in_science	0	740066	2.5488	1.0509	0.0000	7.0000
	1	229908	3.2001	0.7071	1.0000	6.0000
$p < 0.0001$	all	969974	2.7032	1.0188	0.0000	7.0000
gpa_all_courses	0	739696	2.3531	0.6458	0.0000	4.0000
	1	229908	3.0672	0.4884	1.7900	4.0000
$p < 0.0001$	all	969604	2.5224	0.6834	0.0000	4.0000
math_important	0	740066	0.3428	0.4747	0.0000	1.0000
	1	229908	0.3724	0.4835	0.0000	1.0000
$p < 0.0001$	all	969974	0.3498	0.4769	0.0000	1.0000
absorbed_by_math	0	740066	0.3309	0.4705	0.0000	1.0000
	1	229908	0.3601	0.4800	0.0000	1.0000
$p < 0.0001$	all	969974	0.3378	0.4730	0.0000	1.0000
hs_math_reqmnt	0	486619	2.7553	0.6338	1.0000	4.0000
	1	201805	2.8630	0.5619	1.0000	4.0000
$p < 0.0001$	all	688424	2.7869	0.6155	1.0000	4.0000
hs_science_reqmnt	0	489255	2.5517	0.6656	1.0000	4.0000
	1	202134	2.7428	0.5954	1.0000	4.0000
$p < 0.0001$	all	691389	2.6076	0.6517	1.0000	4.0000

Note: n_0 reflects no counterfactual match, n_1 is a counterfactual match.

FIGURE 1. COUNTERFACTUAL IDENTIFICATION FRAMEWORK (BOLD FONT IN PANEL (D) INDICATES DEPARTURES FROM OBSERVABLES, RESULTING IN LOWER WEIGHTS)

(a)

	Four-year degree holders in a STEM discipline (donor pool)												
	Individual academic			Individual attributes		Geographical indicators zip-code				School Level			Outcome
ID	GPA_Math	GPA_Read	GPA_Science	Woman	Ethnicity	Med Income	% poverty	% unemployed	% home owned	% Free/reduced	AP offered	% in AP	Graduate
1	3	4	4	1	White	35357	7	6	50	15	1	35	1
2	2	4	3	1	White	18353	12	10	20	37	0	0	1
3	3	3	4	0	White	35272	6.5	5	45	10	1	25	1
4	4	2	2	0	Af. Amer.	23961	8	7	22	24	0	0	1
5	2	4	3	0	Hispanic	22275	8	7.5	19	25	0	0	1
6	4	3	3	0	Af. Amer.	36415	7	5	55	10	1	20	1
7	3	4	4	1	Af. Amer.	43367	4	3	60	9	1	50	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...
n	3	3.5	4	0	Asian	21209	9	7	21	35	1	18	1

(b) Initial pool of two-year entrants

(c) Counterfactual identification via probability matching

(d)

	Countrfactual two-year entrants identified from the initial cohort after matching											Outcomes		Feasibility weight
	Individual academic			Individual attributes		Geographical indicators zip-code				School Level			Counterfactual or Expected*	(To be) Observed**	Counterfactual probability***
	GPA_Math	GPA_Read	GPA_Sciece	Female	Ethnicity	Med Income	% poverty	% unemployed	% home owned	% Free/reduced	AP offered	% in AP			
	3	4	4	1	White	35357	7	6	50	15	1	35	1	(?) 1	1
	2.2	**3.5**	3	1	White	18353	12	10	20	37	0	0	1	(?) 1	0.95
	3	3	4	0	White	35272	6.5	5	45	**18**	1	25	1	(?) 0	0.99
	4	2	2	0	**Hispanic**	**26961**	8	7	**32**	**10**	**1**	**7**	1	(?) 0	0.81
	2	**3.5**	3	**1**	Hispanic	22275	**2**	7.5	**85**	25	**1**	**35**	1	(?) 0	0.85
	4	3	3	0	Af. Amer.	36415	7	5	55	10	1	20	1	(?) 1	1
	3	4	4	1	Af. Amer.	43367	4	**13**	60	**2**	1	**35**	1	(?) 0	0.91
	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
	3	3.5	4	0	Asian	21209	9	7	21	35	1	18	1	(?) 1	1

*Counterfactual Outcome based on the probabilistic matching.

**This outcome may be compared against outcomes in existing databases (as proposed in this study) or may be collected as they are realized over time (to take place in future projects)

***Feasibility weight refers to the extent to which the success case resembles the characteristics of the counterfactual individual. If all observables are the same, feasibility is 1.

FIGURE 2. PROBABILITIES OF COUNTERFACTUAL IDENTIFICATION NELS SAMPLE

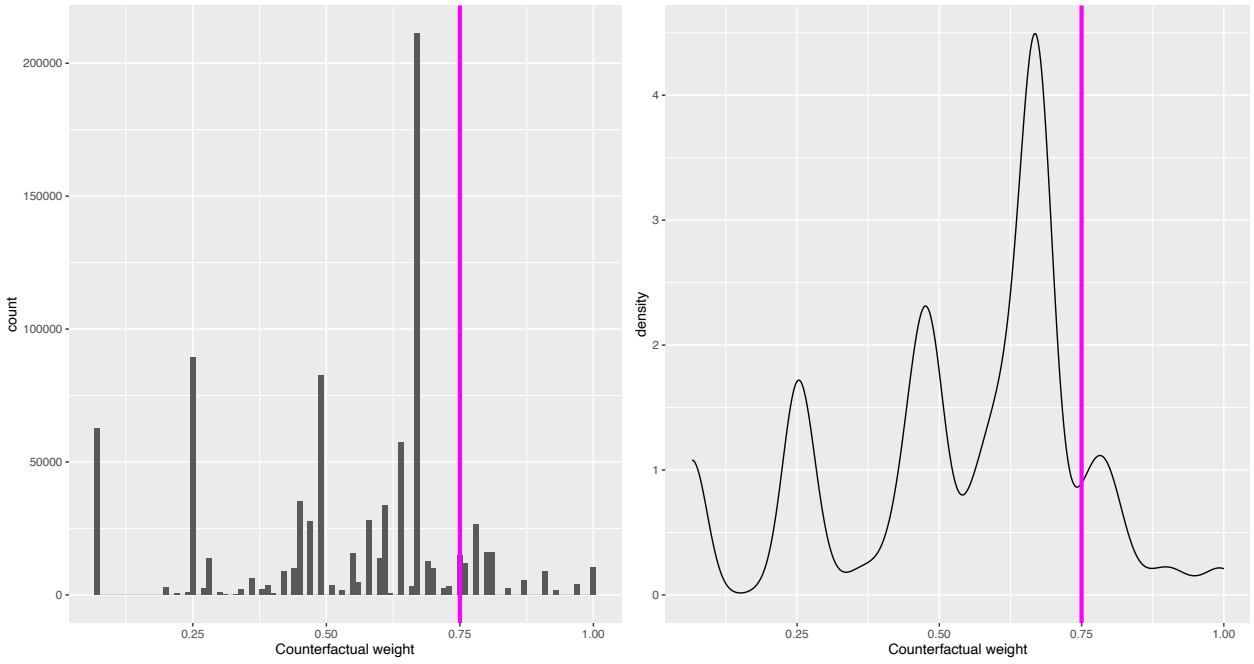


FIGURE 3. PROBABILITIES OF COUNTERFACTUAL IDENTIFICATION ELS SAMPLE

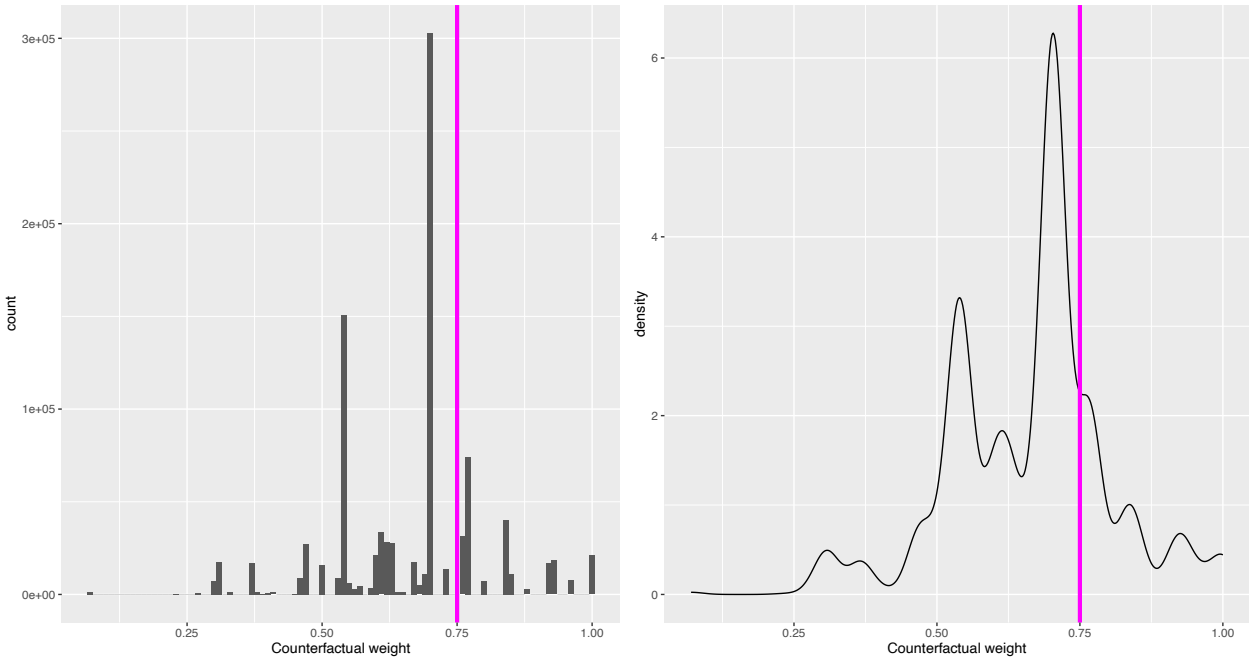


FIGURE 4. BEST PREDICTORS OF SUCCESS IN BS STEM GRADUATION NELS

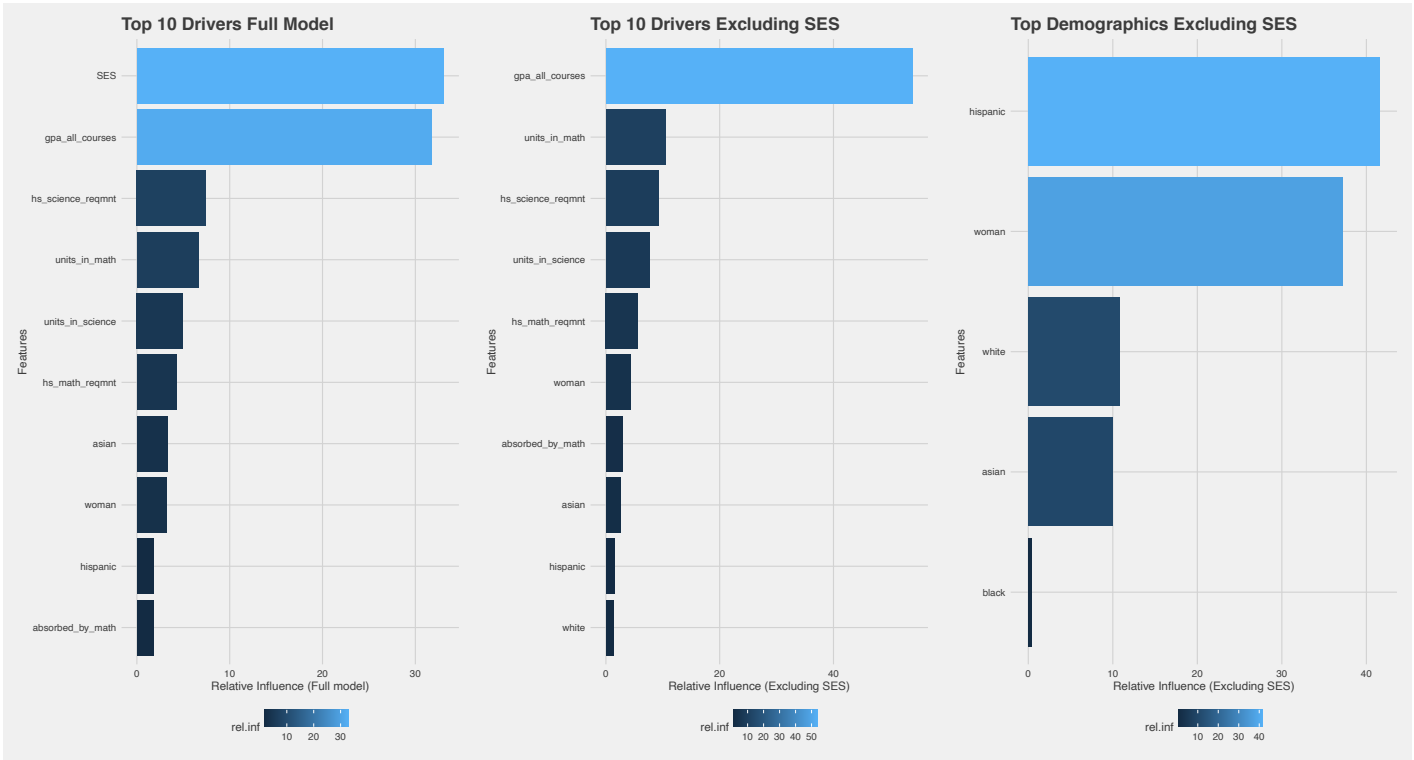
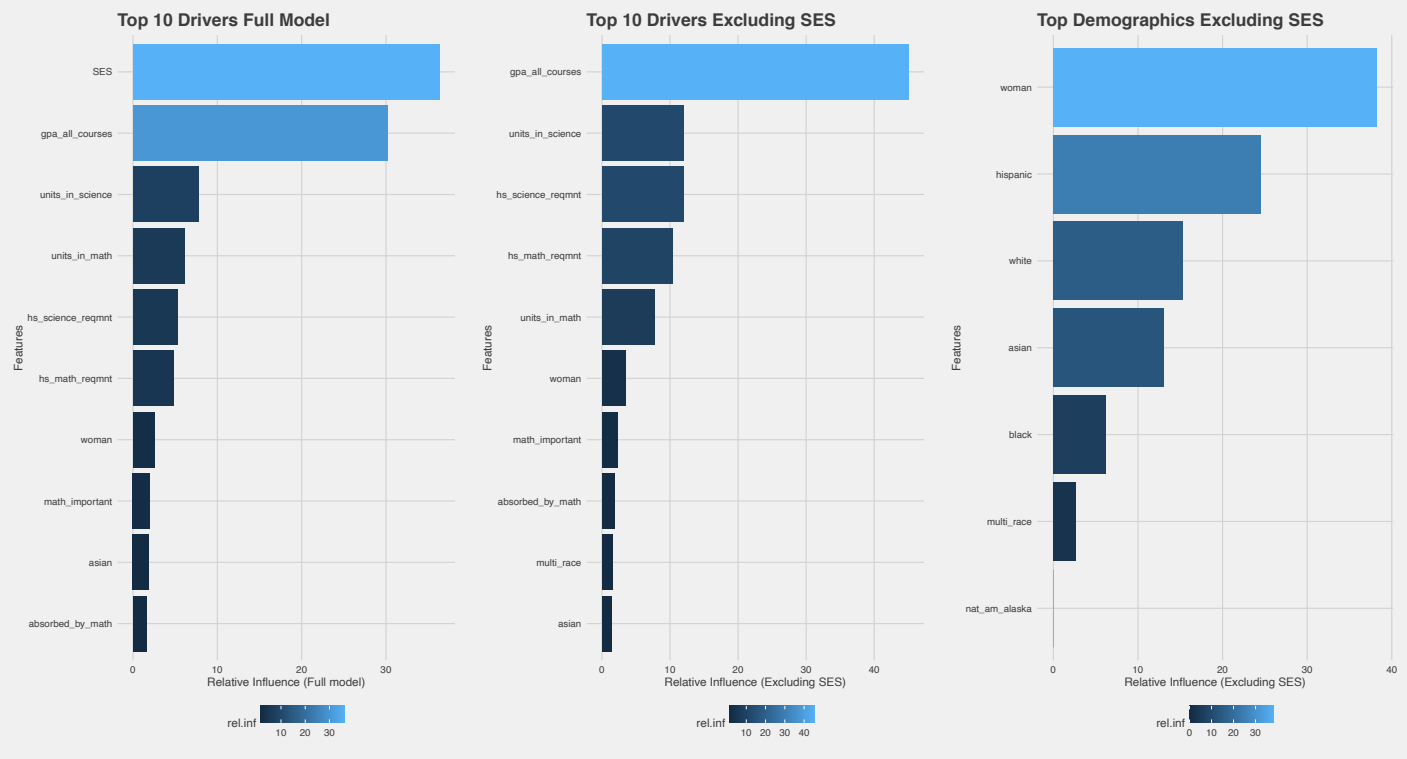


FIGURE 5. BEST PREDICTORS OF SUCCESS IN BS STEM GRADUATION ELS



About the authors

Manuel S. González Canché is a professor at the University of Pennsylvania. His research agenda focuses on developing plans of action to improve the success prospects of students from low-income and minoritized backgrounds. He has assessed the impact of AI algorithms in reproducing inequalities, while he also has relied heavily on data science and interactive visualizations to detect outstanding low-income students who, despite living in places with concentrated disadvantages, were thriving academically. More recently, he has striven to democratize access to data science and interactive visualizations by developing no code free-to-use software for the analysis of qualitative data evidence as well as low-code applications for spatial econometrics .

Chelsea Zhang is currently a PhD candidate in Higher Education at the Graduate School of Education and a master's student in Statistics at Wharton, the University of Pennsylvania. She holds a BS in human development from Kent State University and an MA in higher education from the University of Michigan. Her primary research interests are centered on quantitative aspects of higher education research, including quasi-experimental and mixed methods research for estimating causal impacts and explaining underlying mechanisms through which causal impacts occur. Chelsea is particularly interested in the economic aspects of college access, such as analyzing the impact of college tuition on student enrollment and student loan amounts and in investigating the relationship between research funding and knowledge production, power dynamics, and equity in higher education institutions.

About the TIAA Institute

The TIAA Institute helps advance the ways individuals and institutions plan for financial security and organizational effectiveness. The Institute conducts in-depth research, provides access to a network of thought leaders, and enables those it serves to anticipate trends, plan future strategies, and maximize opportunities for success.

To learn more, visit tiaainstitute.org.

Join the conversation online: @TIAAInstitute

GRE-3757280PR-E0724W

TIAA Institute

(08/24)