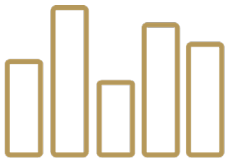


STRATEGIC DATA PROJECT

Understanding patterns of success among postsecondary CTE students: A diagnostic for institutional and system analysts Technical Guide

This diagnostic was prepared with support from the ECMC Foundation. This diagnostic was prepared by Chris Avery, Jon Fullerton, Brian Johnson, Adrienne Murphy, Alyssa Reinhart, and Elise Swanson (listed alphabetically). The authors would also like to thank members of our Advisory Board: Joel McKelvey, Sue Mukherjee, Christopher Leake, and Jessica Cunningham, as well as Miriam Greenberg and the ECMCF SDP CTE Fellows for their feedback and guidance throughout the development of the tool.

Overview of Technical Guide

This technical guide accompanies the main narrative document entitled "Understanding patterns of success among postsecondary CTE students: A diagnostic for institutional and system analysts." The narrative document introduces the Diagnostic: background, goals, terms, overviews of analyses, example visualizations, and suggestions for further reading. Analysts should then refer to this technical guide for more details about how to execute the analyses: data decisions to make, data specifications, model considerations, practice interpreting statistical output, and more.

Key Decision Points: The First Step

Before engaging with the analyses described in this Diagnostic, there are several key decisions that should be made based on the context of your institution or system. **First, decide which student populations you want to analyze.** The students contained in the data analysis file at the time the analysis code is run will be the students represented in the results. If you want to focus on a particular population of students, you should only include that population in the data file. If you want a separate analysis for each of multiple student populations, consider creating a separate data file for each population and running the analysis code on each file. We recommend at least running separate analyses for students enrolled full- or part-time in the entry semester to account for potential differences in expected time to completion, intended credential completion, and other potential differences between these groups. You may want to consider excluding students in dual enrollment for similar reasons.

Second, decide how long of a time horizon you want to consider when examining whether a student has achieved a successful outcome. The time horizon you select should be based on the expected or desired time to completion for pathways included in the analysis. For example, you may want to examine a period of three years after entry for students entering pathways that typically lead to an associate degree and a shorter period for students entering pathways that typically lead to a certificate.

Third, decide which cohorts you want to include in your analysis. Aggregating your analyses across cohorts will increase your sample size for any given analysis. This can provide a technical benefit, like helping to provide sufficient data to fit regression models. Including multiple cohorts can also enable you to see general trends for your institution across time rather than findings driven by some idiosyncrasies of a single cohort. However, if you know you made substantial changes to your programs, curriculum, or student support services in a particular year, you probably do not want to include cohorts from both before and after the changes in your analysis, or you risk obscuring any differences in student outcomes or changes in trends that resulted from those changes.

Fourth, decide how you want to define student pathways. These can be defined based on programs, meta-majors, or some other classification that makes sense in your context. The descriptions we provide are general to allow you to use pathway definition that will allow you to get the most relevant insights for your analysis.

Fifth, decide which pathways you want to compare. We recommend limiting comparisons to a handful of pathways (loosely 2 to 4) at a time to allow you to see differences across pathways and engage in further inquiry based on those differences. This rough limit will ensure that the analysis results are digestible and that the comparisons are helpful. You may want to run through the Diagnostic multiple times using different combinations of pathways to inform discussions with various stakeholders at your institution or across your system.

Finally, decide how you want to define successful outcomes for students. We define it to include instances of completion and transfer, where completion is defined as attainment of any credential and transfer is defined as enrolling in any other institution. You could also define completion more narrowly; for example, earning a credential in a student's initial pathway or receiving the level of credential in which a student first expressed interest. Because completion and transfer are not mutually exclusive, we focus on a student's first successful outcome – but you could also look at whether students ever complete, regardless of whether that occurs after a transfer event. Choose outcome events and definitions that are meaningful for your context.

Once these decisions have been made, you are ready to engage with the Diagnostic. Before starting your analysis, we also recommend acquainting yourself with the statistical software you have available and installing any further programs you may need. Below, we describe the analyses generally so that analysts using a variety of softwares can reproduce these figures. For users of Stata (version 16.1 or later), we provide sample code that should be ready to run with little modification. Certain analyses also draw on Python, a free software you can install on any computer. More information about Python, including instructions for installing it, can be found at https:///www.python.org. We also recommend consulting with your IT department to discuss options for installing and using either of these programs. However, if you do not have access to Stata 16 or Python, you can still engage with the Diagnostic to guide the questions your institution is asking about its career and technical programs and to produce insightful, relevant analyses.

Data

Please refer to the data specifications below for each section for a full account of the data and format required for these analyses. In general, you need access to a longitudinal dataset that follows students from the time of first enrollment at your institution/system. Longitudinal data is necessary to observe completion and transfer events as they occur over time. You also need access to student demographics data, transcript information, and if possible, National Student Clearinghouse (NSC) data to identify transfer and completion outcomes beyond students' initial institution of enrollment.

Note about Example Visualizations

The sections below include example visualizations that are based on synthetic data, not data coming from any actual college or system. The synthetic data were constructed to show patterns similar to those that have been observed in actual data.

Section 1: Data and Analysis Guide

Key Questions	Outcomes of Interest	Key Factors to Consider	Analytic Approach	Limitations
• Are there differences in completion rates across pathways? • Are there differences for students with the same background characteristics?	• Credential completion or transfer **Data Sources:** • Longitudinal student data for multiple cohorts along with demographics and transcript data • National Student Clearinghouse (NSC) records	• Student background characteristics • Full or part-time enrollment in entry term • Relevant time period (e.g., if outcome is associate degree, using at least 3 years of data following entry)	• Multinomial logistic regression – with and without controlling for student background characteristics	• Descriptive, non-causal • Does not capture stacking of credentials

Description of Analyses

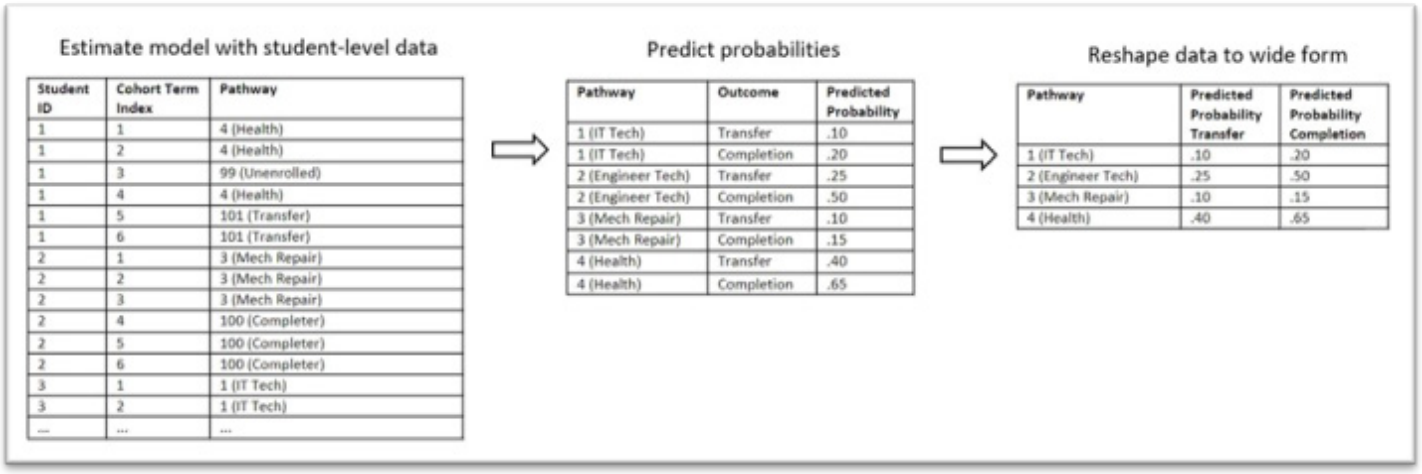
Are there differences in success rates across pathways?

In this analysis, we estimate the probability that a student in a given pathway will have a successful outcome event. We defined success in our analyses to be either completion of a credential of any form (associate degree, certificate, etc.) or transfer to another institution (depending on your context, you may want to specify this as transfer to a four-year institution or transfer to any institution). Please refer to the data specification section for more details about what data you will need to conduct this analysis yourself and in what format to conduct this analysis.

The primary tool for Section 1 is a multinomial logistic regression model, a useful tool when there are multiple levels of the outcome variable. In this case, we have two levels of the outcome measuring student success: credential completion (1) or transfer (2). (These two levels are not inherently ordered; we could instead code transfer (1) and completion (2) without substantively altering the results.) The multinomial logistic regression allows us to estimate separate probabilities for each outcome in a single model, which is more efficient and allows us to account for the competing nature of the two types of success. Students are coded as either transferring or completing based on which event occurred first- for example, if a student completes a certificate and then transfers to a four-year institution, they are coded as a completer, not a transferer.

The first multinomial logistic regression model run expresses students' outcomes only as a function of their initial pathway choice. Once this model has been fit to the data, we use it to predict for each pathway the probability of first completion and first transfer (i.e., the model is fit on student-level data but predicted values calculated at the pathway level). From these predictions, we should obtain a results matrix with one row per pathway-outcome, storing the predictions. For example, if you were analyzing four pathways and looking at transfer and completion for each, you should create a matrix with eight observations. Next, we reshape the results matrix from a long to a wide format, with one row per pathway. For example, if you were analyzing four pathways, your reshaped data would have four observations. Figure 1 illustrates these data transformations.

Figure 1: Data transformation for predicted probabilities charts



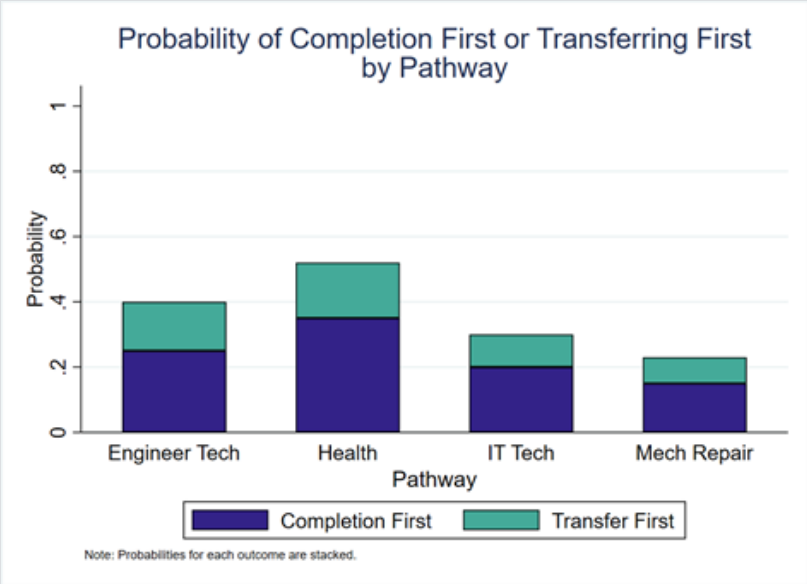
Student ID	Cohort Term Index	Pathway
1	1	4 (Health)
1	2	4 (Health)
1	3	99 (Unenrolled)
1	4	4 (Health)
1	5	101 (Transfer)
1	6	101 (Transfer)
2	1	3 (Mech Repair)
2	2	3 (Mech Repair)
2	3	3 (Mech Repair)
2	4	100 (Completer)
2	5	100 (Completer)
2	6	100 (Completer)
3	1	1 (IT Tech)
3	2	1 (IT Tech)
...	...	...

Pathway	Outcome	Predicted Probability
1 (IT Tech)	Transfer	.10
1 (IT Tech)	Completion	.20
2 (Engineer Tech)	Transfer	.25
2 (Engineer Tech)	Completion	.50
3 (Mech Repair)	Transfer	.10
3 (Mech Repair)	Completion	.15
4 (Health)	Transfer	.40
4 (Health)	Completion	.65

Pathway	Predicted Probability Transfer	Predicted Probability Completion
1 (IT Tech)	.10	.20
2 (Engineer Tech)	.25	.50
3 (Mech Repair)	.10	.15
4 (Health)	.40	.65

Importantly, you do not want to overwrite the original, student-level data used to fit the model to make these transformations, but rather should try to use the software's working memory to create and temporarily store this new dataset of predicted values. If your software does not allow you to store temporary files, be sure to save your results to disk with a new name so that they are accessible for graphing. Whichever approach you take to generate a matrix of predicted results, we use these results to create a stacked bar chart showing the predicted probability of completion and transfer for each pathway. See Example Visualization 1a.

Example Visualization 1a



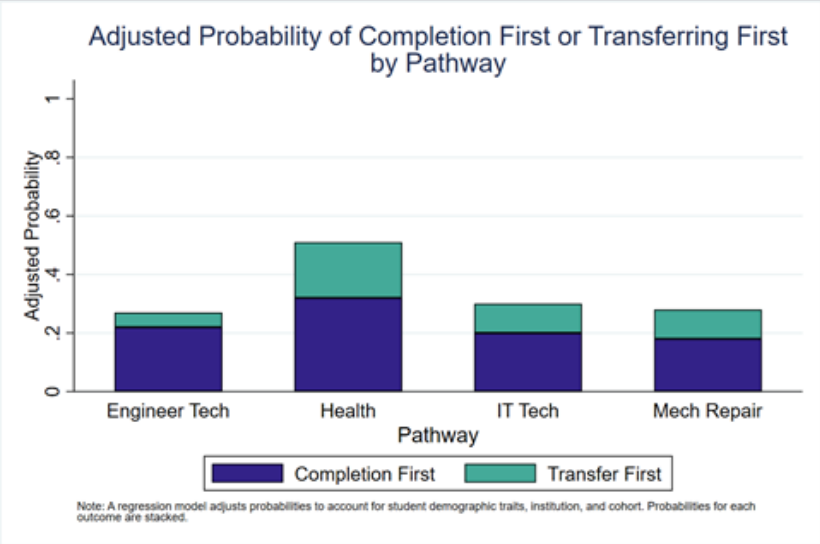
This analysis shows whether students' likelihood of completion or transfer varies depending on what pathway they initially pursue. Critically, this is a descriptive analysis that does not account for student background characteristics, program context, or other factors that influence student outcomes. These results should not be interpreted to mean that one program is outperforming another, but rather to encourage deeper questioning and discussion across pathways.

Are there differences for students with the same background characteristics?

The next stage of this analysis investigates potential differences in student success across pathways after accounting for students' background characteristics. Once again, we use a multinomial logistic regression model, with student outcome (completion or transfer) as the dependent variable. The difference is that now we include several student-level variables, in addition to a student's initial pathway choice. These include gender (measured as a binary male/female), age at enrollment, race/ethnicity (disaggregated to Asian, Black, Latina/o/x, white, and other/unknown), Pell dollars awarded in the entry year, high school GPA, and mother's education level (less than high school, high school, college or more). Finally, we interact each of these additional control variables with a student's initial pathway choice, so that the relationships between student background characteristics and outcomes can vary by pathway.

Like with the first multinomial logistic regression, we fit this model to student-level data, then predict probabilities of completion and transfer for each pathway, for a hypothetical student who has the average of each of the student traits included in the model. These results should again be saved to a pathway-outcome level matrix and transformed into a wide dataset with one pathway per row and each outcome as a separate column. Finally, this dataset should be used to create another stacked bar chart with the adjusted probabilities of successful outcomes by pathway. See Example Visualization 1b.

Example Visualization 1b

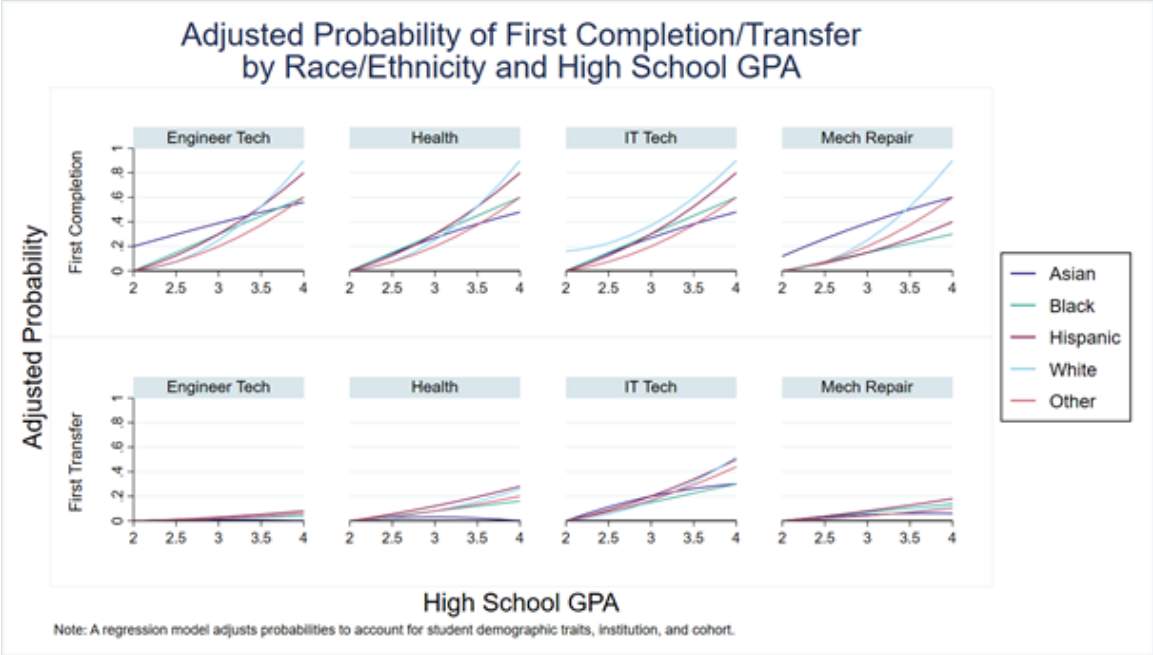


By design, this chart looks similar to Example Visualization 1a. The useful information is in comparing how predicted probabilities of completion or transfer shift – or stay the same – after controlling, or "adjusting," for background student traits. If the predicted probabilities remain largely unchanged in a pathway, this is evidence that student factors in the model – race, gender, Pell dollars awarded, etc. – matter less in whether or not a student is successful. If none of the predicted probabilities shift very much across any of the pathways, this is evidence that differences in student success outcomes are potentially due less to which students select into particular pathways. Instead, differences in success could be due more to pathway-specific factors and structures. If predicted probabilities within a pathway do differ after controlling for student traits, this is evidence that the selection of students into the pathway is explaining some of the observed success (or lack thereof). Compare the Engineer Tech pathway in Example Visualization 1a and 1b. After adjusting for student traits with the regression model, the probability of transfer in particular has dropped substantially. In other words, the regression model is suggesting that the probability of transfer among Engineer Tech students is due largely to students more likely to transfer selecting into Engineer Tech.

We can use the same regression model to further investigate, by pathway, which student factors are most associated with which success outcomes. Specifically, we can select one or two of the student traits included in the model, vary these while holding constant all the other traits, and see how predicted probabilities of success change. For illustration, see Example Visualization 1c. This chart uses the regression model to predict the probabilities of completion and transfer within each pathway, for different racial/ethnic groups, with varying high school GPAs, while holding fixed all other traits in the model. There is much we can learn from such a chart. One conclusion is that across pathway and racial/ethnic groups, a better high school GPA is less predictive of transfer than completion. In other words, across pathways and racial/ethnic groups, holding all other traits fixed, imagine two hypothetical students, one with a very good high school GPA and one with a more typical high school GPA. The student with the higher GPA has a much higher predicted probability of completing a credential than the lower-GPA peer, but a more modest predicted advantage in the probability of transfer compared to the lower-GPA peer.

Many similar charts are possible using results from the same model. Our sample diagnostic code includes charts that use the multinomial logistic regression to explore predicted probabilities by each of gender, high school GPA, age, Pell award, race/ethnicity, and mother's highest level of education, all while holding constant the other variables. We also include examples of charts that explore varying pairwise combinations of these variables while holding all else constant. Feel free to experiment with your own versions of charts, too.

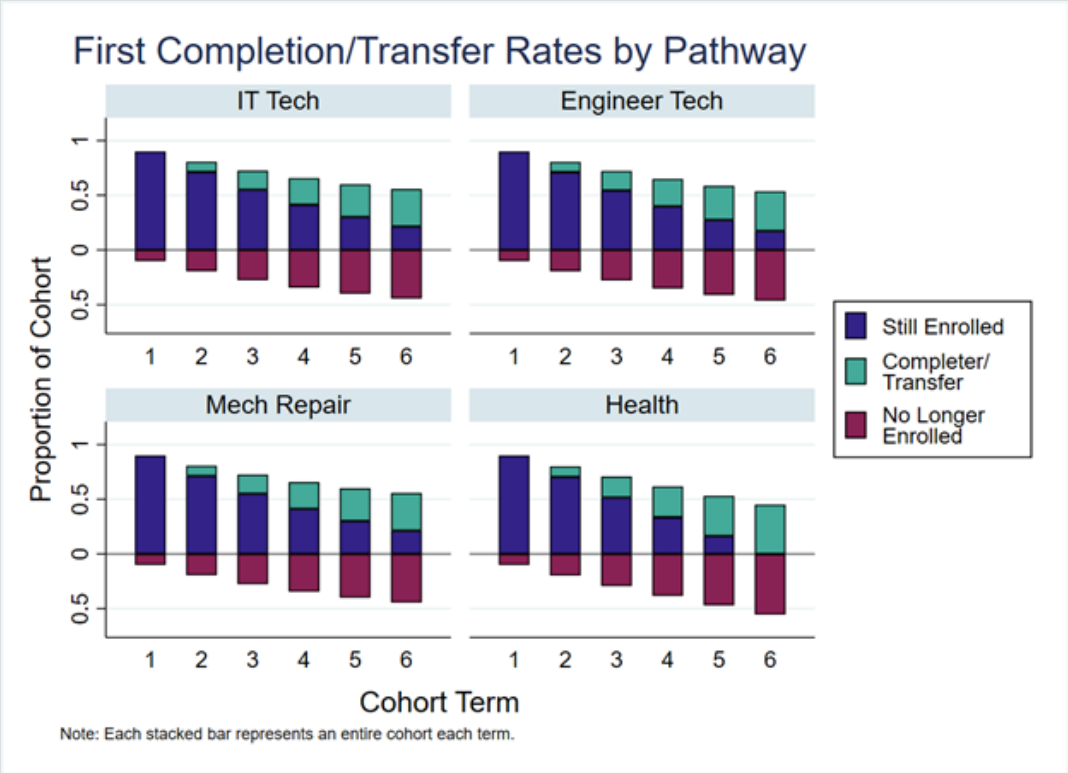
Example Visualization 1c



Are there differences in the timing of success across pathways?

A final chart for Section 1 displays the timing of student success, allowing you to compare not just the extent to which students are successful across pathways but the speed with which they achieve a successful outcome. The chart represents each cohort, in each pathway, in each term, as a single bar. The area of the bar in red below the x axis represents the proportion of students no longer enrolled. Above the x axis, the area of the bar represents the proportion of students still enrolled or those who have completed/transferred, with blue and turquoise distinguishing between the two. As the term is incremented, the graph conveys shifts in the proportion of students in each of the possible states (still enrolled, completer/transfer, no longer enrolled). See Example Visualization 1d.

Example Visualization 1d



Level of Uniqueness

This file should be unique at the studentid-cohorttermindex level. In other words, each row in the data file should represent a unique term of enrollment for a unique student.

Only cohorttermindex and pathway values should vary over time for students in the data set. All other variables should be calculated based on the traits of a student at first entry into your institution and then held fixed for the rest of the terms, however many terms you decide to include.

We need multiple terms of data for each student in order to create the "waterfall" chart (see Example Visualization 1d), which requires observing a student over time. However, the multinomial logistic regression models only require a single observation per student (e.g., the first term, the term our sample code uses) because all the variables used in the model should be constant for a student over time: an outcome measured after however many terms you decide to include in your analysis, an initial pathway choice, and other covariates fixed at entry. In other words, you should not fit the regression models for Section 1 using all the observations in the data set, because each student will have multiple rows. See example Stata code for more information.

Data File Specification

Variable Name	Description	Values	Notes
studentid	Unique student identifier	Numeric	Must be unique to each student
institutionid	Unique institution identifier	Numeric	Must be unique to each institution in the data; most relevant when multiple institutions included in single analysis
cohortyear	Calendar year in which student's cohort first enrolled	Numeric	Can use academic year of entry instead of calendar year; if you want to include more than just fall cohorts, you could consider adding a cohortseason variable, to identify when the cohort entered within the year
cohorttermindex	A value indexing, in order, each term of enrollment for a student	Integer; consecutive integers 1 through max number of terms considered	These indices should be sequential for each student and each student should have the same number of cohorttermindex values, beginning with 1, regardless of entry pathway; if a student unenrolled during a given term, there should still be a row in the data representing that student and term, and for each term thereafter, up through the max number of terms included
pathway	Pathway code identifying a student's pathway affiliation for a given term	Integer; can vary term to term for a student; codes 99, 100, 101 reserved to represent non-enrollment, completion, and transfer, respectively	Pathway code for the record when cohorttermindex equals 1 is a student's entry pathway; codes 100, 101 should be "end state" pathway codes, as in a student remains with that code after first completing or transferring; code 99 need not be considered an "end state" code, in that students might disenroll and then re-enroll, all before completing or transferring

male	Indicator value for being identified as male in the data	Integer; 0 (female), 1 (male)	Encodes student gender as a binary; depending on data availability, sample sizes, and context, could include indicators for additional gender identities
race	Race/ethnicity	Integer; unique value for each race/ethnic category	Race should combine race and ethnicity; Hispanic/Latinx should be one level of race. Your context, data availability, and sample sizes will determine what categories are included
age	Age at entry	Numeric; rounded to nearest year	To calculate, subtract birth year from cohort year
motheredlevel	Mother's highest level of education	Integer; unique value for each education level (e.g., middle school, high school, any college, unknown)	"Unknown" should be a level of this variable, following the FAFSA convention, instead of a separate variable indicating that mother's education level is missing. Depending on your context, you may have a different variable for parental education/first-generation status
pell	Pell grant dollars awarded in academic year of entry	Numeric	Consider replacing missing Pell award values with the mean Pell award value among non-missing observations; including the mi_pell indicator in the regression will still ensure that these missing values are not treated exactly the same way as students who were not missing Pell award
mi_pell	Indicator for missing Pell award information	Integer; 0 (not missing), 1 (missing)	Students missing Pell info. May not have completed FAFSA
hsgpa	High school GPA	Numeric; should be on 0-4.0 scale	Consider replacing missing high school GPA values with the mean high school GPA among non-missing observations; including the mi_hsgpa indicator in the regression will still ensure that these missing values are not treated exactly the same way as students who were not missing GPA. Depending on how GPA is
			reported, you may need to account for weighted and unweighted GPAs
mi_hsgpa	Indicator for missing high school GPA	Integer; 0 (not missing), 1 (missing)	

Notes About Regression Specification

The variables we include in the main regression model for this analysis are not the only factors that affect student outcomes, but they tend to be predictive of outcomes and are commonly recorded by institutions.

Among variables that are relevant and available, we still made choices about what to include and exclude by default for the sake of successfully running a regression. For example, you can see from this data specification document that we ask for mother's highest level of education – but not father's highest level of education. Highest level of education for mother and father are often similar. This strong association makes it doubtful that adding father's highest level of education to the model would contribute much more information, though it would make the regression more demanding on the data to run. However, you could also combine mother's and father's highest level of education into a single variable indicating whether a student is first-generation (i.e., neither parent has a college credential) or operationalize this information in a different way.

If you have other information about your students that you want to include in the main regression, please feel free to do so by adding it to the data and then modifying the regression code appropriately.

Additionally, if you are conducting these analyses for multiple institutions and cohorts of students, we recommend including indicators for students' institution of enrollment and year of matriculation. You can also interact the student-level covariates with the institutional indicators if you think the relationships between student characteristics and outcomes varies by institution.

Notes About Potential Regression Issues

Regression models are helpful because they allow us to control for some set of characteristics, making an "all else equal" comparison with respect to the included variables. Regressions are not without potential technical problems, however.

Be on the lookout for situations in which variables included in the main regression provide the same information. If such a scenario– called "collinearity" – arises, the redundant variables will be automatically dropped from the regression model by most analysis softwares, including Stata. The automatic omission of a variable can then cause problems for calculating margins, which are the predicted values plotted in the graphs for this section.

For example, mother's highest level of education should include a level that represents "unknown." This non-availability might result from a student completing the FAFSA but without providing the requested information about mother's education. Or this non-availability might be a result of a student not completing the FAFSA at all. Imagine a scenario in which every incoming student who completed the FAFSA provided information about mother's education. Then every student whose value of motheredlevel is the integer you

choose to represent "unknown" must not have completed the FAFSA at all. Those same students will not have any Pell grant information, which means they will all have a 1 for mi_pell. Now an indicator variable representing the "unknown" level of motheredlevel and mi_pell provide exactly the same information: identifying students who didn't complete the FAFSA. One of these variables will be automatically omitted from the regression, which can then prevent the analysis software from using the regression model results to do further calculations. You can proactively check whether this is an issue in your data by estimating correlation matrices among your covariates or by reviewing the construction of these variables in your data codebook.

If you know in advance that two or more variables provide the same information, you should select which you want to include in your data and model. This may require you to modify the regression code.

Also, recall from above that the definition of completion you choose may mean more or fewer students qualify as completers in your dataset. Defining completion in terms of finishing any credential would mean more students are coded as completers. Defining completion in terms of finishing the specific credential a student intended to complete upon entry would likely mean fewer coded as completers. Similarly, the time horizon over which you allow students in each cohort to complete may mean more or fewer students coded as completers. With too few completers in the data set, the main regression model might not converge. In that case, the model cannot be used to make predictions and generate charts.

Section 2: Data and Analysis Guide

Key Questions	Outcomes of Interest	Key Factors to Consider	Analytic Approach	Limitations
• When and why do students transfer across pathways and what are the outcomes of these inter-pathway transfers?	• Credential completion or transfer to another institution **Data Sources:** • Longitudinal student data for multiple cohorts w/ demographics and transcript data • National Student Clearinghouse (NSC)	• Student background characteristics • Full or part-time • Relevant time period (e.g., if outcome is associate's degree, using 3 years for time to completion)	• Descriptive analysis	• Descriptive, non-causal

Description of Analysis

When and why do students transfer across pathways and what are the outcomes of transfer?

This analysis produces an interactive Sankey diagram showing flows of students across semesters. We start with a student-term level dataset in which students are grouped according to their initial pathway choice, in their entry term. This dataset is then collapsed to the pathway-transition level to include aggregate counts of the number of students moving into, remaining in, or transitioning out of each possible pathway or outcome (unenrolled, still enrolled in first pathway, transferred to each potential pathway, transferred to another institution, or completed a credential) in each term. See example Stata code for Section 2 for one way to approach collapsing the data in this manner. Regardless of the approach you take, it is important to ensure that a consistent number of students is included in the analysis within an entry pathway, across semester transitions -- for example, after a student completes a credential, they continue to be counted in the "completed" category moving forward rather than being dropped from the analysis. Figure 2 illustrates this transformation of the data.

Figure 2: Data transformation for Sankey diagram

StudentID-CohortTermIndex level data

Student ID	Cohort-Term Index	Pathway
1	1	1 (IT Tech)
1	2	1 (IT Tech)
1	3	1 (IT Tech)
1	4	1 (IT Tech)
1	5	100 (Completer)
1	6	100 (Completer)
2	1	2 (Engineer Tech)
2	2	2 (Engineer Tech)
2	3	7 (Health)
2	4	12 (Culinary)
2	5	99 (Unenrolled)
2	6	99 (Unenrolled)
...	...	...

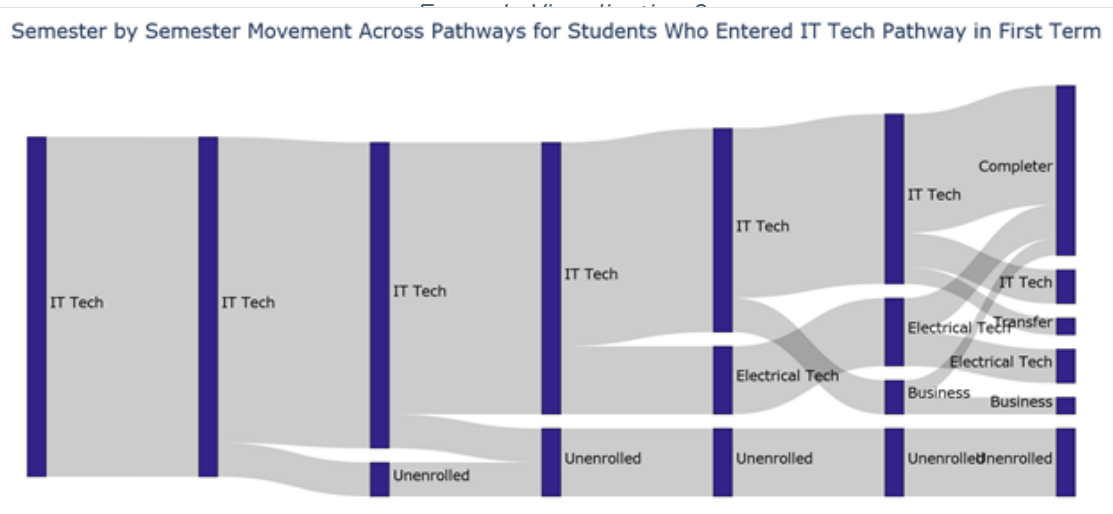
Collapsed data ready for Sankey chart

Entry Pathway	Term Transition	Prior Term Pathway	New Term Pathway	Number Students
1 (IT Tech)	1	1 (IT Tech)	1 (IT Tech)	100
1 (IT Tech)	2	1 (IT Tech)	1 (IT Tech)	90
1 (IT Tech)	2	1 (IT Tech)	99 (Unenrolled)	10
1 (IT Tech)	3	1 (IT Tech)	1 (IT Tech)	80
1 (IT Tech)	3	1 (IT Tech)	99 (Unenrolled)	10
1 (IT Tech)	3	99 (Unenrolled)	99 (Unenrolled)	10
...	...	...	...	...

We recommend setting up your data so that there are two term transitions per year. Depending on the institution's context, this may require rolling summer events into the fall, winter events into the spring semesters, or some other aggregation based on context. This helps smooth out the analysis by limiting the transitions to those relevant to the most students (e.g., at many institutions, fewer students are enrolled over the summer). Or you could include more transitions and code all students who did not enroll for a winter or summer term as remaining in the same pathway, although this may mute patterns of student exit during these terms.

Once the data are set, we use Python to create the Sankey diagrams. See Example Visualization 2a. The diagram displays the size of each flow within each term transition, scaled to the number of students out of the starting total that the flow represents. Using Python enables interactivity such that you can hover your cursor over any particular flow and see the exact number of students making that transition between pathways/outcomes.

While this analysis is useful for showing students' movement across pathways and through outcome events, it is a descriptive, non-causal analysis only. The results should not be interpreted to mean that entering one pathway rather than another causes students to transfer pathways or to ultimately be successful or unsuccessful. Instead, the Sankey charts document students' choices regarding their enrollment over time.



Another limitation of this analysis is that we do not show stacking of credentials or churn between institutions. Once a student has transferred or completed a credential, that classification is carried forward for the remaining terms included in the analysis. Students are coded as "completer" if they earn a credential before transferring and remain with this code for the rest of the analysis, potentially obscuring behavior in which a student earns multiple "stacked" credentials. Similarly, students who transfer to another institution and then re-enroll at your institution are coded as "transfer" as soon as the first transfer event takes place. Such students remain coded as "transfer" for the rest of the analysis, potentially obscuring repeated transferring that is potentially important. However, we do dynamically capture periods of stop out, where a student exits the institution and then re-enrolls; students are not coded as "unenrolled" indefinitely if they return. You should feel free to modify any of these decisions, if another approach would provide more insight in your context.

Level of Uniqueness

This file should be unique at the studentid-cohorttermindex level. In other words, each row in the data file should represent a unique combination of studentid and cohorttermindex.

Data File Specification

Variable Name	Description	Values	Notes
studentid	Unique student identifier	Numeric	Must be unique to each student
cohorttermindex	A value indexing, in order, each term of enrollment for a student	Integer; consecutive integers 1 through max number of terms considered for that entry pathway	Every student associated with a particular pathway at entry should have the same number of terms of enrollment in the data file
pathway	Pathway code identifying a student's pathway affiliation for a given term	Integer; can vary term to term for a student; codes 99, 100, 101 reserved to represent non-enrollment, completion, and transfer, respectively	Pathway code for the record when cohorttermindex equals 1 is a student's entry pathway; codes 100, 101 should be "end state" pathway codes, as in a student remains with that code after first completing or transferring

Note About Student Subpopulations

The students contained in the data analysis file at the time the analysis code is run will be the students represented in the results. If you want to focus on a particular subpopulation of students, you should only include that subpopulation of students in the data file (e.g., just students who were part-time in their first term, just full-time, or all students).

Note About Python

Sankey charts showing student movements across pathways can quickly become visually overwhelming. We chose to set up these charts to plot numeric pathway codes, which are easier to display when many semester-to-semester transitions are shown. Using Python and the plotly graphing package make it possible to display student counts when one hovers the cursor over a given transition. This also saves space. There are downsides to using Python and plotly, however, because the charts must render in a browser to be explored fully. Static images of the charts will not allow this full exploration. We felt the Python-enabled interactivity outweighed these downsides. If you must have static images of charts and do not simply want to take screenshots, there are ways to automatically save out plotly charts as images. See the plotly package documentation for more information.

To make these Sankey charts using the provided code requires that Python and the plotly Python package be installed and accessible to Stata. Python is a free, open-source tool that your IT department should be able to help you set up, if you do not already have it. If your Stata installation has trouble accessing your Python installation, Stata documentation available online can help you troubleshoot this issue. The installation process for plotly can vary depending on how Python has been installed, so this might be another task for which your IT department should be consulted.

Should you desire pathway codes to be strings instead of numeric values (e.g., "Business and Social Sciences" instead of 1) in the Sankey charts, please feel free to alter the provided code, which was set up with only numeric codes in mind. You might consider using abbreviated pathway titles though, to minimize issues with space and overplotting.

At this time, to our knowledge, the plotly package does not offer node sorting options. This means that when the Sankey charts render in the browser, the pathway nodes may appear in a different order from transition to transition. It is also possible that if and when the charts are re-run and re-rendered, the pathway node ordering from transition to transition may be different than what they were for previous chart renderings.

Section 3: Data and Analysis Guide

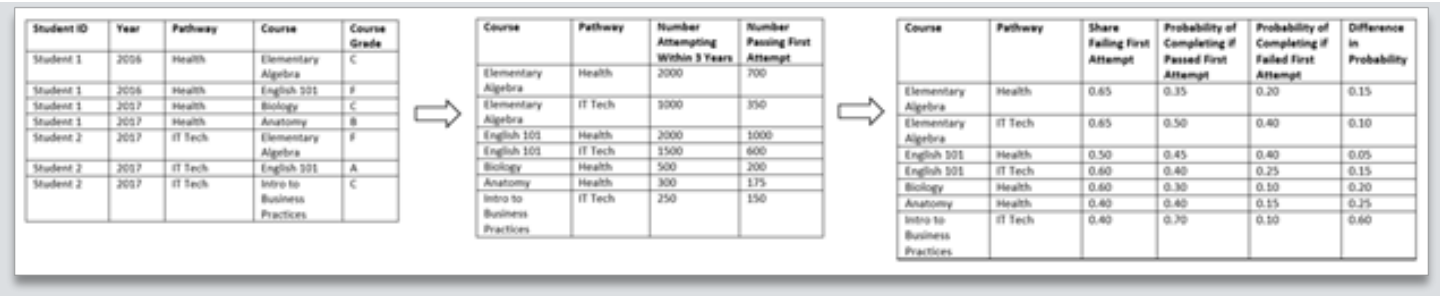
Key Questions	Outcomes of Interest	Key Factors to Consider	Analytic Approach	Limitations
• Why do pathway completion rates differ? Are there gateway courses that are getting in the way of completion for some pathways?	• Credential completion **Data Sources:** • Longitudinal student data for multiple cohorts w/ demographics and transcript data • National Student Clearinghouse (NSC)	• Student background characteristics • Full or part-time • Relevant time period (e.g., if outcome is associate degree, using 3 years for time to completion)	• Descriptive analysis	• Descriptive, non-causal

Description of Analysis

Are there gateway courses that are obstructing completion in some pathways?

This analysis investigates core courses that are required for students to complete within a pathway. These may be general requirements, such as introductory math or English, or pathway-specific requirements. To prepare the necessary data, we start with a student-course-year-level dataset to observe all of the course attempts made by students included in the analysis. We then transform this into a pathway-course level dataset that captures the number of students attempting and succeeding in the required courses within their first three years of enrollment. Next, we use these numbers to calculate the proportion of students who fail each course on the first attempt. Finally, we calculate the probability that students completed a credential within three years of entry, separately among those students who passed or failed the course on the first attempt. To calculate this probability among those who passed (failed), divide the number of students who passed (failed) the course and completed a credential in a pathway by the total number of students who started in that pathway, attempted the course, and passed (failed) that initial attempt. We take the difference of these two probabilities to calculate the difference in the probability of completing a credential between those who initially passed or failed the required course. Figure 3 illustrates this data transformation.

Figure 3: Data Transformations for Gatekeeper Course Analysis



Student ID	Year	Pathway	Course	Course Grade
Student 1	2016	Health	Elementary Algebra	C
Student 1	2016	Health	English 101	F
Student 1	2017	Health	Biology	C
Student 1	2017	Health	Anatomy	B
Student 2	2017	IT Tech	Elementary Algebra	F
Student 2	2017	IT Tech	English 101	A
Student 2	2017	IT Tech	Intro to Business Practices	C

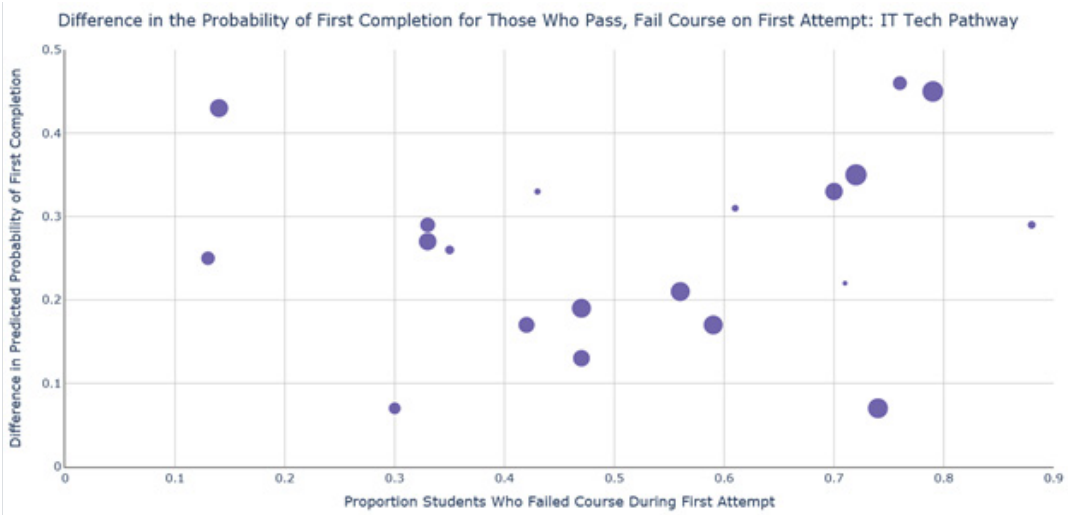
Course	Pathway	Number Attempting Within 3 Years	Number Passing First Attempt
Elementary Algebra	Health	2000	700
Elementary Algebra	IT Tech	1000	350
English 101	Health	2000	1000
English 101	IT Tech	1500	600
Biology	Health	500	200
Anatomy	Health	300	175
Intro to Business Practices	IT Tech	250	150

Course	Pathway	Share Failing First Attempt	Probability of Completing if Passed First Attempt	Probability of Completing if Failed First Attempt	Difference in Probability
Elementary Algebra	Health	0.65	0.35	0.20	0.15
Elementary Algebra	IT Tech	0.65	0.50	0.40	0.10
English 101	Health	0.50	0.45	0.40	0.05
English 101	IT Tech	0.60	0.40	0.25	0.15
Biology	Health	0.60	0.30	0.10	0.20
Anatomy	Health	0.40	0.40	0.15	0.25
Intro to Business Practices	IT Tech	0.40	0.70	0.10	0.60

We then create an interactive scatterplot that plots the share of students failing a course against the change in probability of completing a credential if the course is failed initially. The size of each point reflects the number of students attempting the course. When interpreting this chart, we can think of large points with a high failure rate and a large decrease in probability of completion (e.g., a large point in the top right of the graph) as the most concerning from a student success perspective. See Example Visualization 3a. When these scatterplots are rendered in the browser, you can scroll over an individual point to see the course title, number of attempters, the probability of completion among those who initially pass, and the probability of completion among those who initially fail.

Note that because some courses might not be unique to a single pathway (especially general education courses), courses may appear in multiple scatterplots. The information provided is not necessarily redundant, however, because the same course might play a different gatekeeping role in different pathways.

Example Visualization 3a



Level of Uniqueness

This file should be unique at the pathway-course_id level. In other words, each row in the data file should represent a unique combination of pathway and course.

Data File Specification

Variable Name	Description	Values	Notes
pathway	Code for student's chosen pathway at entry, however pathway is defined for an institution	Integer; unique value for each pathway	
course_id	Unique course identifier that identifies the same course term to term across sections	Integer; unique value for each course included in analysis	Example: course_id 1 might be assigned to Math-101, even though multiple sections of Math-101 might be offered each term
course_label	Course label	String	Example: "Math-101"; will be used for display purposes in scatterplots
total_attempters	Total number of unique students who attempted the course at least once	Integer; should be a positive number	
proportion_failing	Proportion of students failing course on first attempt	Numeric; value should fall between 0, 1	You should decide whether to consider withdrawals failures
prob_completer_fail	Probability of completion first, among students who failed course on first attempt	Numeric; value should fall between 0, 1	Calculation: (# students who completed a credential before dropout or transfer after failing course on first attempt)/(# students who failed course on first attempt)
prob_completer_pass	Probability of completion first, among students who passed course on first attempt	Numeric; value should fall between 0, 1	Calculation: (# students who completed a credential before dropout or transfer after passing course on first attempt)/(# students who passed course on first attempt)
prob_completer_diff	Difference in probability of completion first between those who passed, failed course on first attempt	Numeric; value should fall between negative and positive 1, though values between 0, 1 are more likely	Calculation: prob_completer_diff = prob_completer_pass – prob_completer_fail

Section 4: Data and Analysis Guide

Key Questions	Outcomes of Interest	Key Factors to Consider	Analytic Approach	Limitations
• Why do pathway completion rates differ? Are students attempting and accumulating credits at the pace necessary to earn a credential in a reasonable timeframe? How do students' early experiences shape their success trajectories?	• Cumulative credits earned • Cumulative credits attempted • Credential completion or transfer • Proportion of students no longer enrolled (stopped out) **Data Sources:** • Longitudinal student data for multiple cohorts w/ demographics and transcript data • National Student Clearinghouse (NSC)	• Student background characteristics • Full or part-time • Relevant time period (e.g., if outcome is an associate degree, using 3 years for time to completion)	• Descriptive analysis	• Descriptive, non-causal

Description of Analysis

Why do pathway completion rates differ? How do students accumulate credits and progress in their pathway over time?

We begin this analysis by comparing college-level credit accumulation over time across pathways. The time period included in your data should reflect a reasonable time frame for your context; for example, three years would reflect 150% of expected time to completion of an associate degree. To get the data into a form ready for graphing, we start with a student-term-course-level dataset and calculate, for each term, each student's total number of college-level (excluding developmental) credits earned and attempted. Depending on your institutional context and popularity of summer and/or winter terms, you may want to roll completion events into the fall and spring semesters. We then calculate each pathway's average cumulative college-level credits earned and attempted for each term, sorted with term ascending. Figure 4 shows this data transformation.

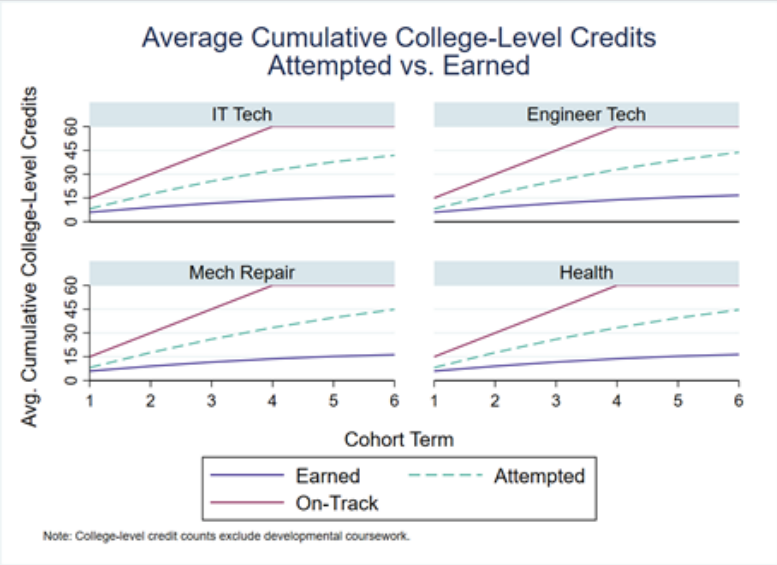
Figure 4: Data transformations for analysis of college-level credits earned vs. attempted

Student ID	Starting Pathway	Semester	College Credits Earned	College Credits Attempted
Student 1	Health	1	12	12
Student 1	Health	…	…	…
Student 1	Health	6	9	9
Student 2	IT Tech	1	3	6
Student 2	IT Tech	…	…	…
Student 2	IT Tech	6	5	8
Student 3	Mech Repair	1	0	0
Student 3	Mech Repair	…	…	…
Student 3	Mech Repair	6	3	3
Student 4	Engineer Tech	1	6	6
Student 4	Engineer Tech	…	…	…
Student 4	Engineer Tech	6	6	6

Pathway	Semester	Average Cumulative College Credits Earned	Average Cumulative College Credits Attempted
Health	1	3.2	6
Health	…	…	…
Health	6	16.9	45.1
IT Tech	1	2.9	5.5
IT Tech	…	…	…
IT Tech	6	15.5	41.8

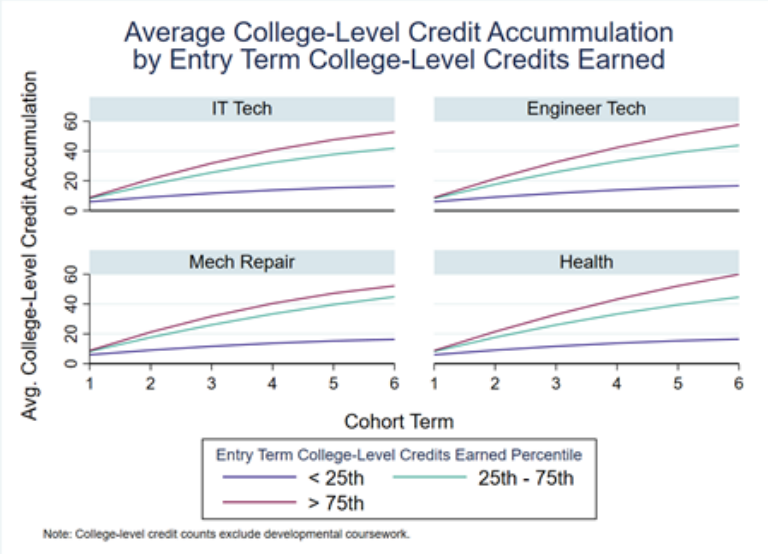
Once you have your data in the form of the second table in Figure 4, you are ready to graph your results in the form of Example Visualization 4a.

Example Visualization 4a



In the second chart, we investigate the importance of early momentum in predicting continued progress. Specifically, we want a chart that plots, by pathway, average college-level credit accumulation among three groups of students: 1) those who were below the 25th percentile in the entry term for college-level credits earned, among other students entering the pathway 2) those who were between the 25th and 75th percentiles and 3) those who were above the 75th percentile. See Example Visualization 4b.

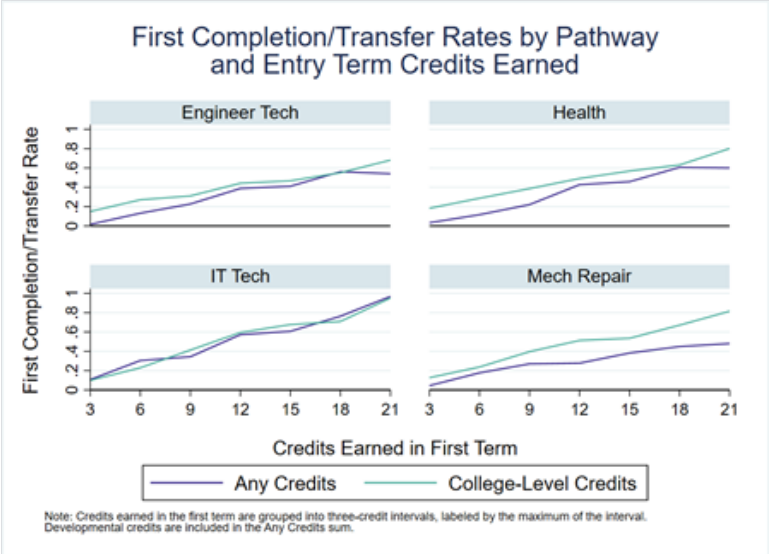
Example Visualization 4b



To get your data into the form required for this graph, you once again should begin with a student-term-course-level data set. Once again, calculate for each student the number of college-level credits earned in each term. Then create a new variable that assigns each student a value ranging from 1-3, depending on whether they fell below the 25th percentile of college credits earned in the first term among pathway peers; between the 25th and 75th percentiles; or above the 75th percentile. This percentile bin value should be constant for a student across terms for the rest of the analysis. Next, collapse the data further to take the average of college-level credits earned by pathway, by term, by percentile bin. Finally, calculate by pathway, term (with terms in ascending order), and percentile bin the cumulative sum of average college-level credits earned. See example Stata code for Section 4 for one possible approach to this data preparation.

In the final graph, we want to investigate the importance of early credit momentum for eventual completion or transfer within three years. Specifically, we want to investigate whether credits earned in the first term predict student completion and transfer rates within three years of entry. Further, we want to know whether college-level credits or any credits (including developmental) differ in how well they predict completion/transfer rates. See Example Visualization 4c.

Example Visualization 4c



For data preparation, start once again with a student-term-course-level data set. From this, calculate for each student, in each pathway, the total number of college-level credits earned in the entry term and the total number of credits earned at any level in the entry term. Next, you will need to create a "coarsened" version of each of these variables. We recommend using bins of 3 credits. For example, if Student A earned 3 college-level credits in the entry term and Student B earned 4.5, consider both students to have earned a college-level credit value falling between 3 and 6; they should both have the same value (3, 6, or something else possibly) for the coarsened version of college-level credits earned. Next, merge on an indicator (a variable with a value of 0 or 1) by Student ID for whether a student completed a credential or transferred within three years of entry. Finally, you will need to collapse your data twice – once for each coarsened measure of college-level credits earned and any credits earned. Each collapse should get you the rate of completion/transfer by pathway, by credit bin value. Append the results of each collapse together while creating an additional indicator for whether a given data row is for college-level credits earned or any credits earned. The final data set for graphing should have columns for pathway, credit bin, completion/transfer rate, and an indicator for credit bin type (college-level credits or any credits). Rows should be unique by pathway, credit bin, and credit bin type. See example Stata code for Section 4 for one possible approach to this data preparation.

As with all the analyses in this Diagnostic, these graphs are descriptive, not causal. They should not be interpreted to mean that, for example, instruction is better in one pathway than another, allowing students in that pathway to accumulate credits at a faster rate. Instead, they should be used to help guide discussions and further inquiry at your institution or system.

Level of Uniqueness

This file should be unique at the studentid-cohorttermindex level. In other words, each row in the data file should represent a unique combination of studentid and cohorttermindex.

Data File Specification

Variable Name	Description	Values	Notes
studentid	Unique student identifier	Numeric	Must be unique to each student
cohorttermindex	A value indexing, in order, each term of enrollment for a student	Integer; consecutive integers 1 through max number of terms considered for that entry pathway	Every student should have the same number of terms of enrollment in the data file, regardless of entry pathway
pathway	Pathway code identifying a student's pathway affiliation for a given term	Integer; can vary term to term for a student; codes 99, 100, 101 reserved to represent non-enrollment, completion, and transfer, respectively	Pathway code for the record when cohorttermindex equals 1 is a student's entry pathway; codes 100, 101 should be "end state" pathway codes, as in a student remains with that code after first completing or transferring
collegecreditsearned	Total college-level credits earned by a student in a term	Numeric; non-negative	
creditsearned	Total credits earned by a student in a term, whether college-level or developmental	Numeric; non-negative	
collegecreditsattempted	Total college-level credits attempted by a student in a term	Numeric; non-negative	