TECHNICAL GUIDE | MAY 2024

Education to Workforce Pathways Diagnostic Toolkit

Analytic code and guidance for using state longitudinal data to understand K12 student progression into postsecondary and the workforce.

ACKNOWLEDGEMENT

The tools and analyses were prepared and supported by Atticus Bolyard, Sue Dynarski, Jon Fullerton, Miriam Greenberg, Julia Lubner, Aleksei Opacic, Eric Taylor, and Rachel Worsham (listed alphabetically).

The authors would also like to thank Elise Swanson, Lisa Sanbonmatsu, Mimi Tan, Alyssa Reinhart, Julia Bloom-Welton, Jayashree Krishnan, Jackie Lundberg, Bonnie Nelson, Dani Fumi, Katie Weaver-Randall, as well as our state partners for their feedback and guidance throughout the development of the tool.

ABOUT THE STRATEGIC DATA PROJECT

SDP partners with state and local K–12 education agencies to build capacity for managing, analyzing, and communicating with data. SDP cultivates analytic talent through a two-year fellowship program, in-person and online trainings, and widely accessible tools and resources. The Harvard Center for Education Policy Research launched SDP in 2008 to meet a need for analytical capacity in state and local K–12 agencies. Reform-minded school superintendents were experimenting with new programs and policies, but lacked the capacity to evaluate those efforts or to make data-informed decisions.

Since 2008, SDP has collaborated with nearly 300 school districts, charter management organizations, state education agencies, and nonprofits to sponsor close to 600 SDP Fellows. Fellows may already work at a partner agency, or SDP recruits and selects Fellows who are then placed at partner organizations. SDP alumni work at K–12 agencies and organizations around the country, and most alumni continue to take advantage of the SDP professional network, trainings, and analytical resources.

The SDP Pathways diagnostic was prepared with support from the Bill and Melinda Gates Foundation.

CONTENTS

INTRODUCTION

This technical guide accompanies the main narrative document entitled "Strategic Data Project Education to Workforce Pathways Diagnostic Toolkit."

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, and more. This guide proceeds as follows:

- First, we discuss key decision points that should be discussed before moving on to analyses. These include decisions like choosing the timeframe in which you will measure outcomes and defining outcomes.
- Second, we provide an overview of the main analytic file, which will be tweaked or reshaped for the subsequent analyses.
- Third, we discuss the file transformations, data needed, and analysis details for each of the visuals presented in the main narrative document. These are organized by section.

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 state.

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. These analyses are centered on high school graduating cohorts from the entire state. Using high school graduating cohorts allows us to compare similarly situated students over time to understand patterns or disparities in student outcomes. However, it is possible that you want to focus on a particular population of students. If so, you should only include that population in the data file.

For example, if you are only interested in students from a certain county or school district, you will need to exclude all other students from the base 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.

Second, decide how long of a time horizon you want to consider when examining students' degree completion and workforce outcomes. The analyses presented in this diagnostic draw on student characteristics and academic achievement as early as 8th grade and trace students' educational and workforce trajectories 10 years from high school graduation.

Given that state longitudinal education data systems (SLEDS) are new, your state may not be able to accommodate analyses that utilize 14 or more years of data. As such, the time horizon you selected should be based on data availability in your state. While examining students' outcomes 10 years or more from high school is optimal, many states also track degree completion and earnings at five and eight years from high school graduation.

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. Including multiple cohorts can enable you to see general trends for your state across time rather than findings driven by some idiosyncrasies of a single cohort. However, if you know your state made substantial changes to secondary or postsecondary education 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 which student characteristics you would like to include in disaggregated analyses. Understanding the extent to which outcomes may vary for students with different academic and demographic characteristics can help your state identify potential strategies for supporting these students.

The disaggregates we explore in this diagnostic are informed by the **Education-to-Workforce Indicator Framework**—a comprehensive guide that includes a common set of metrics and data equity principles for assessing and addressing disparities along the pre-K-to-workforce continuum.[1] We suggest that users leverage this framework to choose disaggregates that fit state needs and interests and

[1] Mathematica (2023) Educator-to-Workforce Framework. https://www.mathematica.org/projects/education-to-workforce-indicator-framework

are easily measurable in your data system.

Finally, decide how you want to define successful outcomes for students. We include college enrollment within one year of high school graduation, completion of a postsecondary credential within 10 years of graduation, and earnings above the living wage benchmark. Based on your knowledge of your state, you could define these outcomes differently. Perhaps you could examine college enrollment within two years or using another more locally relevant wage benchmark.

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 software can reproduce these figures. For users of Stata (version 17.1 or later), we provide sample code that should be ready to run with little modification. However, if you do not have access to Stata 17, you can still engage with the Diagnostic to guide the questions your state is asking about its students' degree completion and workforce outcomes and to produce insightful, relevant analyses.

DATA AND BASIC FILE STRUCTURE

Each section below contains data specifications, which include a full account of the data, definitions, and file format required for the analyses. In general, you need access to a longitudinal dataset that follows students from high school, to college, and through the workforce. Longitudinal data is necessary to observe enrollment, completion, and work events as they occur over time.

You also need access to student demographic data and high school and college transcript information. To fully capture student enrollment and completion behaviors, you will need postsecondary data from both in- and out-of-state institutions. Out-of-state enrollment and completion data can be provided by the National Student Clearinghouse (NSC).

Finally, to complete the wage analyses, you will need access to earnings data for workers in your state. These data are derived from unemployment insurance records and should be linkable to your student records.

The analytic files for most analyses in this Diagnostic will be derived from a main file tracking students over the ten years after high school graduation. This file is long—meaning that a student will have 10 rows of observations each. Each of these rows will contain information about the students' enrollment, employment, and degree completion status in that particular year. Each row will also contain demographic and academic information taken from the students' high school records. If you are unable to track students 10 years from high school graduation due to the length of your panel, consider tracking students five or eight years from high school. See below for important variables and an example of the file structure.

In the variables table, we include the variable name, description, vales, notes, and the name of the variable in the associated Stata .do files (A .do file contains the Stata commands used to complete these analyses). In these tables and the .do file, the variables in all caps must be swapped out for the data element name used in your dataset. For example, your unique student identifier may be named "uniqueid." This will need to be substituted for "STUDENT_ID" in the code that accompanies this Diagnostic. Variables in lower case are generated in the

.do files. We have also included the file format for each chart in a separate excel file titled "Open Diagnostic Data Layout," which can be found in the data_templates folder on the OpenSDP github.

VARIABLES

Variable Name	Description	Values	Notes	Variable in .do
Student ID	Unique student identifier	Numeric	Must be unique to each student.	STUDENT_ID
Cohort Year Index	A value indexing, in order, each year after high school graduation	Integer; consecutive integers 1 through max number of years considered	These indices should be sequential for each student, and each student should have the same number of cohort term index values, beginning with 1 regardless of HS graduation year; if a student was not enrolled and/or not employed during a given year, there should still be a row in the data representing that student and year, and for each year thereafter, up through the max number of years included.	years_from_hs
HS Graduation Year	Academic year in which a student's cohort graduated	Numeric	Can either use the fall or spring term year.	HS_GRADUATION_YEAR
Enrolled indicator	Indicator noting whether a student was enrolled in college for at least 1 semester during the academic year	Indicator/ integer	Analysts will need to consider how to classify summers. In these analyses, we include preceding summers as a part of an academic year. Ex. Summer 2012 is included in the 2012-13 school year. Analysts should also understand which institutions are covered in their SLEDS. Ideally, the SLEDS should include in-state public and private colleges, as well as institutions out-of-state (provided by the National Student Clearinghouse).	IN_COLLEGE_INDICATOR
Working indicator	Indicator noting whether a student was present in the labor data in the year measured	Indicator/ integer	Analysts should note that not all employment is covered by UI wage data. Those missing include: those who are working out of state and those whose position is not eligible for UI benefits (self-employed, federal employment).	WORKING

Variable Name	Description	Values	Notes	Variable in .do
Highest credential	Indicator noting the students' highest credential earned in or before that academic year	Indicator/ integer	This indicator will change over time if a student earns other degrees. As with enrollment data, the analyst should understand which colleges are and are not included in their data system. To fully capture the outcomes of a state's students, the SLE DS will need to include completion information from in- and out-of-state colleges.	HIGHEST_DEGREE_IN_YEAR
Demographics & academic characteristics	Demographics and academic characteristics of interest (race, socioeconomic status, sex, HS test scores, etc.)	Indicator/ integer	Likely to be indicators or categorical variables.	SEX, RACE, etc.
Degree in HS	Indicator variable for whether a student received a postsecondary credential in high school	Indicator/ integer		DEG_IN_HS

SHAPE OF DATA

Data should be shaped to include one row per student per year after high school graduation.

Student ID	HS Graduation Year	Cohort Year Index	Enrolled	Working	Highest credential	Sex
001	2012	1	1	0	HS/GED	Female
001	2012	2	1	0	HS/GED	Female
001	2012	3	1	1	HS/GED	Female
001	2012	4	1	1	Bachelor's	Female
001	2012	5	0	1	Bachelor's	Female
001	2012	6	0	1	Bachelor's	Female
001	2012	7	0	1	Bachelor's	Female
001	2012	8	0	1	Bachelor's	Female
001	2012	9	1	0	Master's	Female
001	2012	10	0	1	Master's	Female

LEVEL OF UNIQUENESS

This file should be unique at the student id-cohort year index level. In other words, each row in the data file should represent a unique year after high school graduation for a unique student.

Section 1: Patterns in Educational Attainment

Data and Analysis Guide

Section one examines patterns in degree completion 10 years from high school graduation to establish a foundational understanding of students' educational outcomes after they depart from high school.

Below, we describe the variables needed for these analyses, the shape of the data, and level of uniqueness needed in the file.

NOTE ON NUMBERING

The numbering of the Diagnostic Charts in this technical guide matches the numbering of the charts in the main narrative document. For example, if you would like to recreate Diagnostic Chart 1.A.1, which shows the percentage of students from a given high school graduating class by highest degree 1-10 years from graduation, you will need to follow the guidance shown below for Diagnostic Chart 1.A.1.

DATA FILE SPECIFICATION FOR DIAGNOSTIC CHARTS 1.A.1-1.A.3:

Associated .do file: "Section_1_College_Completion.do"

Variable Name	Description	Values	Notes	Variable in .do
Student ID	Unique student identifier	Numeric	Must be unique to each student.	STUDENT_ID
Cohort Year Index	A value indexing, in order, each year after high school graduation	Integer; consecutive integers 1 through max number of years considered	These indices should be sequential for each student, and each student should have the same number of cohort term index values, beginning with 1 regardless of HS graduation year; if a student was not enrolled and/or not employed during a given year, there should still be a row in the data representing that student and year, and for each year thereafter, up through the max number of years included.	years_from_hs
HS Graduation Year	Academic year in which a student's cohort graduated	Numeric	Can either use the fall or spring term year.	HS_GRADUATION_YEAR

Variable Name	Description	Values	Notes	Variable in .do
Credential completed indicator	Indicator noting whether a student completed a post-secondary credential in that academic year or any prior academic year	Indicator/ integer	This indicator will remain "on" once a student earns a credential. If they earn further credentials, this indicator will not change.	hasdegree
Highest credential	Indicator noting the students' highest credential earned in or before that academic year	Indicator/ integer	This indicator will change over time if a student earns other degrees.	HIGHEST_ DEGREE_IN_ YEAR
Demo-graphics	Demographics of interest (race, socio-economic status, sex, etc.)	Indicator/ integer	Likely to be indicators or categorical variables.	RACE, SEX, etc.
Pre-college academic perfor-mance	Standardized measures of academic performance prior to college	Integer/ categorical	Likely to be a standardized test score. If students are able to take tests in different years, analysts should consider standardizing scores within test year. We recommend binning students into quartiles.	MATH_TEST_ QUARTILE

SHAPE OF DATA FOR DIAGNOSTIC CHART 1.A.1

Data should be limited to include one row per student per year after high school. These rows should include information captured 1-10 years after high school graduation

.

Student ID	HS Graduation Year	Cohort Year Index	Highest credential
001	2012	1	HS/GED
001	2012	2	HS/GED
001	2012	3	HS/GED
001	2012	4	Bachelor's
001	2012	5	Bachelor's
001	2012	6	Bachelor's
001	2012	7	Bachelor's
001	2012	8	Bachelor's
001	2012	9	Master's
001	2012	10	Master's

SHAPE OF DATA FOR DIAGNOSTIC CHARTS 1.A.2 AND 1.A.3

Data should be limited to include one row per student that includes information captured 10 years after high school graduation.

Student ID	HS Graduation Year	Cohort Year Index	Completed Credential	Highest Credential	Race/Ethnicity	Test score quartile
001	2012	10	1	Bachelor's	Black	4
002	2013	10	0	HS/GED	Hispanic	2

LEVEL OF UNIQUENESS

This file should be unique at the student id-cohort year index level. In other words, each row in the data file should represent a unique year after HS graduation for a unique student.

ANALYSIS SUMMARIES

Analytic Technique Diagnostic Chart 1.A.1: Calculate the percentage of students with each degree type in each year after high school graduation.

Analytic Technique Diagnostic Chart 1.A.2: Calculate the percentage of students with each degree type by demographic characteristics at the 10th year after high school graduation.

Analytic Technique Diagnostic Chart 1.A.3: Calculate the percentage of students in each demographic and test quartile group who have earned a college credential within 10 years of high school graduation.

SAMPLE RESTRICTIONS:

Sample Restrictions for Diagnostic Chart 1.A.1:

- Keep students in high school graduation cohorts you can observe in postsecondary data 1-10 years after graduation.
- Include only graduates who received regular or advanced diplomas (i.e., exclude students who received special education (SPED) diplomas and other certificates).

Sample Restrictions for Diagnostic Charts 1.A.2-1.A.3:

- Keep students in high school graduation cohorts you can observe in postsecondary data 1-10 years after graduation.
- Keep the 10th year after HS graduation.
- Include only graduates who received regular or advanced diplomas (i.e., exclude students who received SPED diplomas and other certificates).

NOTES ON DEMOGRAPHIC VARIABLES

In the Diagnostic Toolkit, we draw demographic and academic characteristics from across a student's high school enrollment. Many of these variables (race, sex, scores from tests taken at the end of high school) can be applied to the base cohort of high school graduates, which is used in most of the analyses in this Diagnostic.

However, if you wish to complete analyses using predictors of enrollment from early in a student's high school career—9th grade GPA or absenteeism, for example—you may wish to change the cohort for those analyses from high school graduating classes to cohorts of 9th graders. This sample is likely more appropriate than simply merging these characteristics onto your graduate file because 9th grade performance is highly predictive of high school graduation. By using a cohort of graduates for these outcomes, you will likely censor your data at the lower end—meaning, you will drop many observations who performed poorly in 9th grade and did not graduate. Using 9th grade cohorts will allow you to fully capture performance in 9th grade.

DATA FILE SPECIFICATION FOR DIAGNOSTIC CHART 1.A.4:

Associated .do file: “Section_1_College_Completion.do”

VARIABLES

Variable Name	Description	Values	Notes	Variable in .do
Student ID	Unique student identifier	Numeric	Must be unique to each student.	STUDENT_ID
Cohort Year Index	A value indexing, in order, each year after high school graduation	Integer; consecutive integers 1 through max number of years considered	These indices should be sequential for each student, and each student should have the same number of cohort term index values, beginning with 1 regardless of high school graduation year; if a student was not enrolled and/or not employed during a given year, there should still be a row in the data representing that student and year, and for each year thereafter, up through the max number of years included.	years_from_hs
HS Graduation Year	Academic year in which a student's cohort graduated	Numeric	Can either use the fall or spring term year.	HS_GRADUATION_YEAR
Credential completed indicator	Indicator noting whether a student completed a postsecondary credential in that academic year or any prior academic year	Indicator/integer	This indicator will remain “on” once a student earns a credential. If they earn further credentials, this indicator will not change.	hasdegree
Stop out indicator	Indicator noting whether a student enrolled in college for at least 1 semester, left college for at least 1 year, and has not earned degree at the time when stop out is observed.	Indicator/integer	The user's definition of stop out can vary based on their local context. This indicator should remain “on” even if the student re-enrolled.	stopout
Currently enrolled, no degree indicator	An indicator for whether the student is currently enrolled in the year measured and does not have a degree.	Indicator/integer	When collapsing the data, analysts will need to ensure that those counted as currently enrolled, no degree do not in fact have a degree.	enrolled

Variable Name	Description	Values	Notes	Variable in .do
Entered college indicator	Indicator for whether a student ever enrolled in college.	Indicator/ integer	This indicator should remain "on" even if the student stopped out.	everenrolled
Degree in HS	Indicator variable for whether a student received a postsecondary credential in high school	Indicator		DEG_IN_HS

SHAPE OF DATA

Data for the Sankey diagram should be collapsed so that the analyst captures students' paths within the 10 years after high school. The transition from one state (e.g., graduated from high school to entered college) to the other is indicated by the source and destination variables. The layer variable orders the sources and destinations. We provide code in the .do file to help analysts collapse their data to this format.

The flow of this diagram from left to right can be thought of as tracing student activities in the 10 years after high school graduation. In prior analyses in this section, we restrict outcomes to a certain time period (degree attainment 10 years after high school).

For the intermediate outcomes in this chart (enrolling in college and stopping out), we do not perform such restrictions. Rather, we are capturing whether a student enrolled or stopped out within 10 years of graduation. Conversely for the terminal outcomes (completed, still enrolled, not enrolled), we measure this state in the 10th year after high school graduation.

ORIGINAL ANALYTIC FILE

Student ID	HS Graduation Year	Cohort Year Index	Currently enrolled, no degree indicator	Stop out indicator	Credential completed indicator	Ever enrolled indicator	Highest credential
001	2012	1	1	0	0	1	HS/GED
001	2012	2	0	1	0	1	HS/GED
001	2012	3	1	1	0	1	HS/GED
001	2012	4	1	1	1	1	Bachelor's
001	2012	5	0	1	1	1	Bachelor's
001	2012	6	0	1	1	1	Bachelor's
001	2012	7	0	1	1	1	Bachelor's
001	2012	8	0	1	1	1	Bachelor's
001	2012	9	1	1	1	1	Master's
001	2012	10	0	1	1	1	Master's

TRANSFORMED FILE

Source	Destination	Layer	Percent of population
Graduated from HS	Entered College	0	75
Graduated from HS	Never Entered College	0	25
Entered College	Never Stopped Out	1	30
Entered College	Stopped Out	1	45
Never Stopped Out	Enrolled	2	5
Never Stopped Out	Completed	2	25
Stopped Out	Not Enrolled	2	30
Stopped Out	Enrolled	2	5
Stopped Out	Completed	2	10

LEVEL OF UNIQUENESS

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

ANALYSIS SUMMARIES

Analytic Technique Diagnostic Chart 1.A.4:
First, calculate the percentage of students who entered college within 10 years of high school graduation.

Then calculate the percentage of students who entered college and stopped out, as defined by leaving college for at least 1 year. Next, calculate the percentage who stopped out and were not enrolled, were enrolled, or had completed a credential in year 10. Finally, calculate the percentage who never stopped out and were enrolled or had completed a credential in year 10.

SAMPLE RESTRICTIONS

- Keep students in high school graduation cohorts you can observe in postsecondary data 1-10 years after graduation.
- Include only graduates who received regular or advanced diplomas (i.e., exclude students who received SPED diplomas and other certificates).
- Keep the 1st –10th years after graduation.

Section 2: K12 to College Enrollment

Data and Analysis Guide

Section 2 analyzes patterns in college enrollment within one year of high school graduation to better understand drivers of credential completion rates. We focus on this timeframe, as secondary schools and districts have the most leverage to impact college-going in this timeframe.

DATA FILE SPECIFICATION FOR DIAGNOSTIC CHARTS:

Associated .do file: “Section_2_College_Going.do”

VARIABLES

Variable Name	Description	Values	Notes	Variable in .do
Student ID	Unique student identifier	Numeric	Must be unique to each student.	STUDENT_ID
Cohort Year Index	A value indexing, in order, each year after high school graduation	Integer; consecutive integers 1 through max number of years considered	These indices should be se-quential for each student, and each student should have the same number of cohort term index val-ues, beginning with 1 regardless of HS graduation year; if a student was not enrolled and/or not em-ployed during a given year, there should still be a row in the data representing that student and year, and for each year thereafter, up through the max number of years included.	years_from_hs
HS Graduation Year	Academic year in which a student's cohort graduated	Numeric	Can either use the fall or spring term year.	HS_GRADUATION_YEAR
Enrolled indicator	Indicator noting whether a student was enrolled in college for at least 1 semester during the academic year	Indicator/integer	Analysts will need to consider how to classify sum-mers. In these analyses, we in-clude preceding summers as a part of an academic year. Ex. Summer 2012 is included in the 2012-13 school year.	IN_COLLEGE_INDICATOR

Variable Name	Description	Values	Notes	Variable in .do
College sector	Sector of college where student enrolled	Indicator/integer	Analysts will need to consider how to deal with stu-dents who enroll in more than 1 college in a given year. For these analyses, we only included students who enrolled in 1 college which was 95% of our sample. If you plan to look at enrollment by other institutional char-acteristics, you will need to include those in this file, as well.	SECTOR
Demographics	Demographics of interest (race, socio-economic status, sex, etc.)	Indicator/integer	Likely to be indicators or categorical variables.	RACE, SEX, etc
Pre-college aca-demic per-for-mance	Standardized measures of academic performance prior to college	Integer/ categorical	Likely to be a standardized test score. If students are able to take tests in different years, analysts should consider standardizing scores within test year. We recommend binning students into quartiles.	MATH_ TEST_ QUARTILE

SHAPE OF DATA FOR CHART 2.A.1

Data should be limited to include one row per student per year after high school. The rows should include information captured 1-5 years after high school graduation.

Student ID	HS Graduation Year	Cohort Year Index	Enrolled
001	2012	1	0
001	2012	2	0
001	2012	3	1
001	2012	4	1
001	2012	5	1

SHAPE OF DATA FOR CHART 2.A.2-2.B.2

Data should be limited to include one row per student that includes information captured one year after high school graduation.

Student ID	HS Graduation Year	Cohort Year Index	Enrolled	College Sector	Race/Ethnicity	Test score quartile
001	2012	1	0	NA	Asian	3
002	2013	1	1	Public 4-year	Black	2

LEVEL OF UNIQUENESS

This file should be unique at the student id-cohort year index level. In other words, each row in the data file should represent a unique year after high school graduation for a unique student.

ANALYSIS SUMMARIES

Analytic Technique Diagnostic Chart 2.A.1: Calculate the percentage of students who enroll in college by time elapsed since graduation.

Analytic Technique Diagnostic Chart 2.A.2: Calculate the percentage of students who enroll in college within one year of high school graduation by college system.

Analytic Technique Diagnostic Chart 2.B.1: Calculate the percentage of students who enroll in college within one year of high school graduation by demographic characteristics.

Analytic Technique Diagnostic Chart 2.B.2: Calculate the percentage of students in each demographic and test quartile group who have entered college within one year of high school graduation.

SAMPLE RESTRICTIONS

Sample Restrictions for Diagnostic Chart 2.A.1:

- Keep students in high school graduation cohorts you can observe in postsecondary data 1-5 years after graduation.
- Keep the 1st-5th years after HS graduation.

Sample Restrictions for Diagnostic Chart 2.A.2

- Keep students in high school graduation cohorts you can observe in postsecondary data one year after graduation.
- Keep the 1st year after high school graduation.
- Keep students who enrolled in college.
- Include only graduates who received regular or advanced diplomas (i.e., exclude students who received SPED diplomas and other certificates).

Sample Restrictions for Diagnostic Charts 2.B.1 & 2.B.2

- Keep students in high school graduation cohorts you can observe in postsecondary data one year after graduation.
- Keep the 1st year after high school graduation.
- Include only graduates who received regular or advanced diplomas (i.e., exclude students who received SPED diplomas and other certificates).

Section 3: College Completion and Stop Out

Data and Analysis Guide

Section 3 tracks college completion patterns among high school graduates with a particular emphasis on enrollment behaviors and composition of college stop outs. The results of these analyses provide further context to the analysis of potential drivers of degree attainment rates in your state.

For these analyses, we define a **stop out** as a student who:

1. Was enrolled in college for at least 1 semester.
2. Did not have a college degree prior to their first term of enrollment.
3. Did not complete a college degree at the time completion was measured
4. And was not enrolled at the time degree completion was measured.

This definition is adapted from work by the National Student Clearinghouse. You may identify stop outs differently based on your local context. Unlike sections 1 and 2, the analyses for stop out require different data transformations. We provide information about the file format below.

DATA FILE SPECIFICATION FOR DIAGNOSTIC CHARTS:

Associated .do files: "Section_3_Stop_Out.do"

VARIABLES

Variable Name	Description	Values	Notes	Variable in .do
Student ID	Unique student identifier	Numeric	Must be unique to each student.	STUDENT_ID
HS Graduation Year	Academic year in which a student's cohort graduated	Numeric	Can either use the fall or spring term year.	HS_GRADUATION_YEAR
Credential completed indicator	Indicator noting whether a student completed a postsecondary credential at the timepoint measured.	Indicator/ integer	The stop out, cur-rently enrolled, and credential completed indi-cators should be mutually exclu-sive. In this diag-nostic, we meas-ure enrollment status within 3 or 6 years of post-secondary entry, depending on the analysis.	DEGREE_ COMPLETER

Variable Name	Description	Values	Notes	Variable in .do
Stop out indicator	Indicator noting whether a student enrolled in college for at least 1 semester, is not currently enrolled and does not have a degree at the timepoint measured.	Indicator/ integer	The user's defini-tion of stop out can vary based on their local con-text. The stop out, currently enrolled, and credential completed indicators should be mutually exclusive. In this diagnostic, we measure enrollment status within 3 or 6 years of postsecondary entry, depending on the analysis.	STOPOUT
Currently en-rolled, no degree indicator	An indicator for whether the student is currently enrolled in the year measured and does not have a degree.	Indicator/ integer	When collapsing the data, analysts will need to ensure that those counted as currently enrolled, no degree do not in fact have a degree. The stop out, currently enrolled, and credential completed indicators should be mutually exclusive. In this diagnostic, we measure enrollment status within 3 or 6 years of postsecondary entry, de-pending on the analysis.	CURRENTLY_ENROLLED
Analysis inclusion indicator	An indicator for whether a student entered col-lege in time to have their outcomes measured at 3 or 6 years post-entry.	Indicator/ integer	For most analyses, this indicator would note whether a student entered college in time for you to measure their enrollment status six years after entry. For example, if a student entered college in fall 2012 and your panel includes data in or past the summer 2018 semester, this individu-al would receive a "1".	OUTCOME_MEASURED
Degree in HS	Indicator variable for whether a student received a postsecondary credential in high school	Indicator/ integer		DEG_IN_HS
Number of hours earned in first term of enrollment	Number of hours earned in first term of enrollment	Numeric	Analysts will likely not have access to transcript data for out-of-state institutions, therefore these analyses may be limited to colleges in the state.	CREDIT_HOURS_FIRST_TERM

Variable Name	Description	Values	Notes	Variable in .do
GPA in the first term of enroll-ment	GPA in first term of enrollment	Numeric	Different institutions may have different conven-tions for calculating GPA. We recommend that analysts calculate their own GPA based on a standardized scale. Analysts will likely not have access to transcript data for out-of-state institutions, therefore these analyses may be limited to colleges in the state.	GPA_FIRST_TERM
Number of hours earned prior to first stop out	Number of hours earned prior to first stop out	Numeric	This variable sums the number of hours earned over the entirety of a students' time in college prior to the first instance of stop out. Analysts can modify this based on their knowledge of how credits were applied to degree programs. Ana-lysts will likely not have access to transcript data for out-of-state institutions, therefore these analyses may be limited to colleg-es in the state.	CREDIT_HOURS_CUM
First term	Indicator noting a students' first term of en-roll-ment.	Indicator/integer	This indicator is the first term a student was enrolled in college ever, not the first term a student enrolled at the institution from which they eventually stopped out.	FIRST_TERM
Term prior to first stop out	Indicator noting the term in which a student stopped out for the first time.	Indicator/integer	The user's definition of stop out can vary based on their local context.	STOP_OUT_TERM
Attempted cre-dential in first term	Indicator noting the credential a student was at-tempting in each school year.	Indicator/integer	For this analysis, we will preserve the credential attempted in the student's first term.	ATTEMPT-ED_DEGREE_FIRST_TERM
Attempted cre-dential in term before stop out	Indicator noting the credential a student was at-tempting in the term prior to their first stop out.	Indicator/integer	For this analysis, we will preserve the credential attempted in the student's last term before stop out.	ATTEMPT-ED_DEGREE_STOP_OUT

Variable Name	Description	Values	Notes	Variable in .do
Re-enrolled indicator	An indicator for whether the student stopped out and re-enrolled	Indicator/ integer		EVER_REEN-ROLLED
Re-enrolled term	Indicator noting the term a student re-entered college after their first stop out.	Indicator/ integer	Our analysis only examines re-enrollment after first stop out; however, you may adjust your analysis to include all instances of stop out and re-enrollment.	REENROLL_ SEMESTER
Time elapsed between stop out and re-enrollment	An integer containing the count of years (or semesters) that passed between stop out and re-enrollment	Integer	In our example, we use years between stop out and re-enrollment; however, this is flexi-ble.	time_be-tween
Demographics	Demographics of interest (race, SES, sex, etc.)	Indicator/ integer	Likely to be indicators or categorical variables.	RACE, SEX, etc
Pre-college academic performance	Standardized measures of academic performance prior to college	Integer/categorical	Likely to be a standardized test score. If students are able to take tests in different years, analysts should consider standardizing scores within test year. We recom-mend binning students into quartiles.	MATH_TEST_ QUARTILE

SHAPE OF DATA

Data should be limited to include one row per student that includes the information above. The symbol "." indicates that data should be missing.

Student ID	Analysis inclusion indicator	Stop out indicator	Degree completer	Currently enrolled, no degree indicator	Entry term	Term prior to first stop out	Re-enrolled term	Time elapsed between stop out and re-enrollment
001	1	1	0	0	Spring 2014	Fall 2017	Fall 2019	2 years
002	1	0	1	0	Summer 2011	.	.	.
003	0	.	.	.	Fall 2012	.	.	.

LEVEL OF UNIQUENESS

This file should be unique at the student id level. In other words, each row in the data file should represent a unique student.

ANALYSIS SUMMARIES

Analytic Technique Diagnostic Chart 3.A.1: Calculate the percentage of postsecondary enrollees who completed a credential, stopped out, or are still currently enrolled at 6 years after college entry.

Analytic Technique Diagnostic Chart 3.A.2: Calculate the percentage of students in each demographic and test quartile group who have entered college and stopped out within 6 years of college entry.

Analytic Technique Diagnostic Chart 3.A.3: Calculate the percentage of postsecondary attempters within degree type who stop out within 3 or 6 years of initial college entry by credits earned or GPA in the first semester.

Analytic Technique Diagnostic Chart 3.A.4: Calculate the number of credits a student was away from the minimum required credits for their attempted credential. Bin credits into 15-hour intervals and plot the percentage of stop outs who fall into each category.

Analytic Technique Diagnostic Chart 3.B.1: Calculate percentage of stop outs who return by duration of stop out.

SAMPLE RESTRICTIONS

Sample Restrictions for Charts 3.A.1 & 3.A.2:

- Keep students for whom you can observe college enrollment data for six years after college entry. We have chosen six years, as it represents 150% of the time it should take to complete a bachelor's degree. For example, if your panel is 10 years long, you will need to limit the sample to students who entered college in or before the 4th year after high school graduation. You can change this based on your sample.
- Include only graduates who received regular or advanced diplomas (i.e., exclude students who received SPED diplomas and other certificates).
- Keep students who attempted college.
- Keep the 6th year after college entry.

Sample Restrictions for Diagnostic Chart 3.A.3:

- Keep students in high school graduation cohorts you can observe completing college three (for sub-baccalaureate degree attempters) or six (for BA attempters) years after college entry. We have chosen six years for BA attempters, as it represents 150% of the time it should take to complete a bachelor's degree. The same logic applies to sub-baccalaureate attempters. For example, if your panel is 10 years long, you will need to limit the sample for BA attempters to students who entered college in or before the 4th year after high school graduation. You can change this based on your sample.
- Include only graduates who received regular or advanced diplomas (i.e., exclude students who received SPED diplomas and other certificates).
- Keep students who attempted college but had not earned a postsecondary credential in high school.
- Keep students for whom you have transcript data.
- Keep the three or six years after college entry.

Sample Restrictions for Diagnostic Chart 3.A.4:

- Keep students in high school graduation cohorts you can observe completing college three (for sub-baccalaureate degree attempters) or six (for BA attempters) years after college entry. We have chosen six years for BA attempters, as it represents 150% of the time it should take to complete a bachelor's degree. The same logic applies to sub-baccalaureate attempters. For example, if your panel is 10 years long, you will need to limit the sample for BA attempters to students who entered college in or before the 4th year after high school graduation. You can change this based on your sample.

- Include only graduates who received regular or advanced diplomas (i.e., exclude students who received SPED diplomas and other certificates).
- Keep students who stopped out of college and had not earned a postsecondary credential in high school.
- Keep students for whom you have transcript data throughout the entirety of their enrollment.
- Keep the three or six years after college entry

Sample Restrictions for Diagnostic Chart 3.B.1:

- Keep students for whom you can observe college enrollment data for six years after college entry. We have chosen six years, as it represents 150% of the time it should take to complete a Bachelor's degree. For example, if your panel is 10 years long, you will need to limit the sample to students who entered college in or before the 4th year after high school graduation. You can change this based on your sample.
- Include only graduates who received regular or advanced diplomas (i.e., exclude students who received SPED diplomas and other certificates).
- Keep students who stopped out of college and had not earned a postsecondary credential in high school.
- This analysis examines re-entry after stop out, therefore you need to allow students the same amount of time to re-enter. In our analyses, we limit the sample to students who stopped out with at least five years in your panel to re-enroll. For example, if your panel is 10 years long, you will need to drop stop outs whose last semester enrolled before stop out was in or after year five post-HS. Depending on the length of your panel, you may consider allowing students three years to re-enter instead of five.

Section 4: Earning a Living Wage

Data and Analysis Guide

Section 4 presents wages for high school graduates across educational attainment and student background characteristics, such as student poverty and prior achievement. We provide context to these analyses by benchmarking wages against living wage thresholds to better understand who is earning enough to meet basic needs.

DATA FILE SPECIFICATION FOR DIAGNOSTIC CHARTS:

Associated .do file: "Section_4_Earnings.do"

VARIABLES

Variable Name	Description	Values	Notes	Variable in .do
Student ID	Unique student identifier	Numeric	Must be unique to each student.	STUDENT_ID
Cohort Year Index	A value indexing, in order, each year after high school graduation	Integer; consecutive integers 1 through max number of years considered	These indices should be sequential for each student, and each student should have the same number of cohort term index values, beginning with 1 regardless of HS graduation year; if a student was not enrolled and/or not employed during a given year, there should still be a row in the data representing that student and year, and for each year thereafter, up through the max number of years included.	years_from_hs
HS Graduation Year	Academic year in which a student's cohort graduated	Numeric	Can either use the fall or spring term year.	

Variable Name	Description	Values	Notes	Variable in .do
Credential completed indicator	Indicator noting whether a student completed a postsecondary credential in that academic year or any prior academic year	Indicator/integer	This indicator will remain "on" once a student earns a credential. If they earn further credentials, this indicator will not change.	
Highest credential	Indicator noting the students' highest credential earned in or be-fore that academic year	Indicator/integer	This indicator will change over time if a student earns other degrees.	HIGHEST_DEGREE_IN_YEAR
Adjusted wages	Annual wages for a student adjusted for inflation.	Numeric	Wages should be adjusted for inflation to the most recent year in the panel. Analysts should consider how to deal with missing wage data. We include guidance below.	ADJ_ANNUAL_WAGES
Demographics	Demographics of interest (race, socioeconomic status, sex, etc.)	Indicator/integer	Likely to be indicators or categorical variables.	RACE, SEX
Program of Study	CIP code for program of study for highest degree at the time wages were measures.	Indicator/integer	Analysts should consider aggregating CIP codes up to broad categories, such as CTE vs liberal arts or broad fields such as education and health sciences. See Dynarski et al. for an example.	PROGRAM_OF_STUDY
Living wage benchmark	The local wage rate that a full-time worker requires to cover the costs of their family's basic needs where they live.	Numeric	The wage benchmarks should be drawn from the year in which you are measuring data and adjusted to the same year as your wage data. See notes below on recommendations for benchmarks.	

SHAPE OF DATA

Data should be limited to include one row per student that includes information captured 10 years after high school graduation.

Student ID	HS Graduation Year	Cohort Year Index	Highest Credential	Adjusted Annual Wages	Race/Ethnicity
001	2012	10	Bachelor's	$50,000	Asian
002	2013	10	Certificate	$30,000	Hispanic

LEVEL OF UNIQUENESS

This file should be unique at the student id-cohort year index level. In other words, each row in the data file should represent a unique year after high school graduation for a unique student.

ANALYSIS SUMMARIES

Analytic Technique Diagnostic Chart 4.A.1: Calculate mean earnings for students with three or more wage quarters present 10 years from high school by highest credential earned.

Analytic Technique Diagnostic Chart 4.A.2: Generate bins for wages (here we create 10 bins), then calculate the percentage of earners who fall into each bin.

Analytic Technique Diagnostic Chart 4.A.3: Plot mean wages by highest credential, test score quartile, and demographic/academic characteristics.

SAMPLE RESTRICTIONS

- Keep students in high school graduation cohorts you can observe in wage data 10 years after graduation.
- Include only graduates who received regular or advanced diplomas (i.e., exclude students who received SPED diplomas and other certificates).
- Drop those without three or four quarters of positive earnings 10 years out from high school. We explain this procedure below, but you may change this based on practices in your state.
- Keep the 10th year after high school graduation.

NOTES ON WAGE DATA

The wage data housed in most state P20W systems are derived from unemployment insurance (UI) data. While these data cover most employed individuals in the state, there are several key sources of missingness that will affect estimates of the returns to college. UI data only include those working in the state and do not include federal and self-employment. As such, we are unable to distinguish between unemployed persons and those who have left the state for other employment, are federal employees (military), or are self-employed. As such, earnings estimates generated by analysis using state UI wage data may be subject to bias due to systematic differences in rates of migration, unemployment, labor force participation, and selection into non-covered employment.

If unemployment is higher among people with only a high school diploma, we may be underestimating the wage differential between high school and college. If people with more advanced degrees are more likely to move out of state and those who move out of state are likely to be higher earners, excluding the out-of-state earners (zero earnings in state) may be further underestimating the wage differential.

Academic work on this topic generally takes one of two strategies for dealing with missing wage data. Most studies focus on workers with in-state earnings (Andrew, Li, Lovenheim, 2016[1]; Altonji and Zimmerman 2018[2]; Andrews and Stange, 2019[3]), dropping workers with no in-state earnings over some time frame. Other papers retain non-matched workers, setting their earnings to zero, often in conjunction with a bounding exercise (Denning, Marx, Turner, 2018[4]). Lee (2009[5]) proposed a bounding approach to estimate treatment effects in the presence of sample attrition. Bounding exercises are likely to be inappropriate in our context because educational attainment likely effects both migration and likelihood of any employment. States that report wages for their students often follow the former example—dropping students without positive wages in the time measured. In addition to contending with completely missing wages in a given time period, we must also consider how to approach individuals who

[1] Andrews, R, J. Li, and M. Lovenheim, 2016. "Quantile Treatment Effects of College Quality on Earnings. Journal of Human Resources, 51(1): 201-238.

[2] Altonji, J. and S. Zimmerman 2019. The Costs of and Net Returns to College Major. In C. Hoxby and K. Stange eds., Productivity in Higher Education, Chicago, Illinois: University of Chicago Press.

[3] Andrews, Rodney J., and Kevin M. Stange. 2019. "Price Regulation, Price Discrimination, and Equality of Opportunity in Higher Education: Evidence from Texas." American Economic Journal: Economic Policy, 11 (4): 31-65.

[4] Denning, Jeffrey T., Benjamin M. Marx, and Lesley J. Turner. 2019. "ProPelled: The Effects of Grants on Graduation, Earnings, and Welfare." American Economic Journal: Applied Economics, 11 (3): 193-224.

[5] Lee, D., 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," Review of Economic Studies 76: 1071–1102.

are missing some, but not all quarters of wage data in a year. Prior work generally either drops individuals with any missing wage quarters in a year or drops those with more than one wage quarter and imputes the missing quarter.

In our work completing this diagnostic with state partners, we used the following procedure:

1. Dropped individuals with more than one missing wage quarter in a year.
2. Linearly imputed the missing quarter using data from other wage quarters in a given year (described below).
3. Dropped observations with imputed wage values below the 1st or above the 99th percentiles (with dollar values of X and Y, respectively).

To complete linear imputation, you will need to follow the procedure we describe below (see code in Section_4_Earnings.do for a demonstration). For individuals with three out of four wage quarters observed 10 years after high school graduation, we linearly impute the missing wage quarter as follows: for each quarter of earnings in the 10th year after high school graduation, we fit a linear model for each quarter of earnings on the three other quarters, among individuals who had wage information observed in all four quarters. This results in a set of four models: one for quarter 1, one for quarter 2, and so on.

For an individual missing a given wage quarter, we predict the missing quarter using that individual's three observed wage quarters, based on the model for that quarter. For example, the first model is fit to predict quarter 1 earnings among individuals who worked consistently 10 years after high school graduation. If an individual is missing wage information from quarter 1, then we use this first model as well as the individual's earnings in quarters 2 through 4 to predict their quarter 1 earnings.

This procedure has several drawbacks. Since we are unable to observe wages of workers who move out of state or are in non-covered employment, we are likely underestimating the returns to bachelor's degrees and higher, as these individuals are more likely to migrate out of state for higher paying jobs. Further, those without college credentials are more likely to not participate in the labor force. As such, by dropping individuals with fewer than three quarters of wage data rather than imputing zeros for these observations, we are likely overestimating the returns of a high school diploma. See Foote & Stange (2022)[6] for a complete discussion of bias that arises from using UI data to estimate the returns to postsecondary education.

[6] Foote, A. & Stange, K.M. (2022) Attrition from administrative data: Problems and solutions with an application to postsecondary education (National Bureau of Economic Research Working Paper No. 30232) http://www.nber.org/papers/w30232

NOTES ON WAGE THRESHOLDS

The living wage benchmark in this diagnostic is taken from MIT's living wage calculator.

For our analyses, we used the living wage for one adult in the year in which we measured wages (e.g., 2021 for the HS class of 2011). If you plan on using multiple cohorts for your analyses, you will need to ensure that the living wage benchmarks for each cohort are inflation adjusted to a common year.

You may also consider benchmarking against other metrics, such as the University of Washington Self-Sufficiency Standard or locally relevant measures. You may also wish to incorporate benchmarks that measure the economic value that degrees provide to students relative to earnings among their state population.

The Postsecondary Value Commission offers four measures and associated documentation that facilitate this kind of comparison.

- Threshold 0 compares wages to a measure of Minimum Economic Return, which is median state-level high school earnings plus total net price amortized over 10 years.
- Threshold 1 compares wages to median earnings for a specific credential level within a state.
- Threshold 2 compares wages to the median earnings of advantaged peers (white, male for credential level within state).
- Finally, Threshold 4, Economic Mobility, compares wages to the 4th (60th to 80th percentile) income quintile or above regardless of credential level within state.

RELATED RESOURCES

- Strategic Data Project Education to Workforce Pathways Diagnostic Toolkit
- Strategic Data Project Education to Workforce Pathways Analysis Narrative
- Strategic Data Project Education to Workforce Pathways Overview
- Strategic Data Project Education to Workforce Pathways GitHub
- Strategic Data Project Tools and Resources

STRATEGIC DATA PROJECT

Harvard's Strategic Data Project works with education agencies to find and train data leaders to uncover trends, measure solutions, and effectively communicate evidence to stakeholders. Our inspiring network of system leaders, fellows, and faculty come together to share how to best use data to make a difference in the lives of students.

Learn more at sdp.cepr.harvard.edu.

Center for Education Policy Research
HARVARD UNIVERSITY