The PNPI Explorer Methodology Report

Purpose of the Explorer

The PNPI Explorer provides the policymaking community with key data points in an easily consumable format. Whether users are congressional staff, advocates, researchers, policy students, reporters, journalists, or state-based officials, the Explorer democratizes access to key postsecondary data points by enabling easier access to and visualization of state and congressional district analyses.

Using data from four federal sources, the PNPI Explorer enables users to create fully customizable higher education profiles at the national, state, U.S. territory, and congressional district levels, distinguishing our tool from other available options. The Explorer aggregates institutional-level data into easily readable charts and graphs. Users can examine how a given state, congressional district, or U.S. territory (geography) intersects with variables such as enrollment, race, degrees conferred, and student debt; compare data with state and national averages; and see how variables have changed over time at the district, territory, state, and national levels. For every analysis, users can generate reports that provide simple, clear, and visually appealing state-by-state and district-by-district comparisons for all available metrics.

The PNPI Explorer was developed with longevity and timeliness in mind. The data infrastructure was built in Stata, creating an easily manageable workflow that requires minimal effort to incorporate new data as they are released. To promote transparency and reproducibility, the Stata code used to generate the data and analyses is publicly available.

Institution Inclusion/Exclusion Criteria

Institutions in the PNPI Explorer are limited to those available in the Integrated Postsecondary Education Data System (IPEDS), the College Scorecard, and the Federal Student Aid (FSA) Data Center.

Additionally, the following types of institutions were excluded from the PNPI Explorer data pull:

- Institutions listed in the FSA Data Center as being a campus outside of the United States;
- Institutions that did not match the IPEDS geographic characteristics (they had no linking FIPS);
- Institutions that were outside of the Office of Postsecondary Education (OPE) universe;
- Institutions that were exclusively online (and thus did not link to a geography);
- Institutions in territories that are not exclusively affiliated with the United States (Marshall Islands and Federated Republic of Micronesia);
- Institutions that had 66% or more of their undergraduate enrollment exclusively online. (After plotting the distribution of undergraduate enrollment exclusively online, we found

that the number of institutions tapers off considerably at the 66% threshold, leading us to use that as our cut point. Histograms of these data are available within the provided code.);
- Institutions that were not degree-granting; and
- Institutions that did not participate in federal Title IV programs

Geographic Considerations

Our analysis is based on institutions' geographic location rather than their students' current residences. Because most of our data come from IPEDS, the College Scorecard, and the FSA Data Center, the geography we are discussing refers to the congressional district, state, or territory in which the institution is located. When displaying the share of students in the two-year repayment cohort who are in default, these are students who attended institutions in the given state, congressional district, or territory. That does not necessarily mean that those students still reside in the congressional district, state, or territory in question. While we limit the institutions in our data to those without large shares of online enrollment, online students who attend other institutions that offer online learning are still included in these data. Therefore, the median student debt of an institution may, in some cases, include individuals who do not live in the same congressional district, state, or territory as the institution.

A second geographic limitation arises from how IPEDS structures its data reporting, which affects the accuracy of our matches with FSA Data Center and College Scorecard data. IPEDS requires institutions to report data according to a Program Participation Agreement (PPA); specifically, the institution's reporting agent reports data for the campus where the PPA is held. There is considerable variation in how this requirement is applied. Some institutions have multiple campuses that report their data separately; thus, each campus (and geography) has campus- and geography-specific data reported in IPEDS. Others report data for all branches at the main campus, as is the case at Ivy Tech Community College in Indiana. Despite Ivy Tech having locations across Indiana, IPEDS reports its data as all in Indianapolis, making geographic aggregation imprecise.

Most Recent Congressional Districts

Our data use the most recent congressional districts available during our data year as the geographic basis of analysis. Congressional district data are available in IPEDS, and have been updated to reflect the boundaries of the 119th Congress. To validate the congressional districts reported in IPEDS, we use the latitude and longitude provided in IPEDS for each institution and run it through the U.S. Census Bureau's TIGERweb API, which geocodes addresses and provides district information. We cross-reference these new districts with the prior districts provided in IPEDS, highlighting any discrepancies. Among those locations where the new district differs from the most recent year, we manually checked the correct district using the Census Bureau's My Congressional District look-up feature.

For the Trend Explorer longitudinal data, we assign the current congressional district for all years included in the data.

Data Sources & Variables

Data for the PNPI Explorer come from the four central public data sources outlined above. Below are the variable categories from each source. For a complete list of variable names, labels, and corresponding aggregate metrics, please refer to the codebook.

Data Source	Variables
IPEDS	**1. Congressional District ID** **2. State FIPS & Name** **3. Institutional Sector (collapsed)** a. For-Profit b. Non-Profit c. Public 2-Year d. Public 4-Year **4. Fall Enrollment by Race** a. Full-Time Undergraduate b. Part-Time Undergraduate c. Full-Time Equivalent (FTE) d. Total Undergraduate e. Total Graduate **5. Fall Enrollment by Gender** a. Full-Time Equivalent (FTE) **6. Fall Enrollment by Age Group** a. Age 24 & Under FTE b. Age 25 & Up FTE **7. Fall Enrollment by Distance Education** a. Split by “None”, “Some”, and “Exclusive” **8. Degrees Conferred by Race** a. Associate b. Bachelor c. Graduate (Master and Doctorate Combined) **9. Degrees Conferred by Gender** a. Men FTE b. Women FTE **10. Graduation Rates at 150% Time** **11. Tuition & Fees by Residency** a. In-District b. In-State c. Out-of-State **12. Tuition Discount Rate** **13. Room, Board, & Other Fees** a. On-campus & Off-campus estimates **14. State Appropriations** **15. Institutional Grant Aid**
College Scorecard	**1. Cost of Attendance** **2. Net Price by Family Income**

	a. All students (in cohort) b. Less than $30K c. More than $110K **3. 3-Year Cohort Default Rate** **4. Median Student Debt** a. By Family Income b. By Completion Status **5. 2-Year Repayment Status for All Undergraduates** a. Making Progress b. Not Making Progress c. Delinquent d. In Deferment e. In Forbearance f. Paid-in-Full g. Fully Discharged h. In Default **6. 2-Year Repayment Status for Undergraduate Non-Completers** a. Making Progress b. Not Making Progress c. Delinquent d. In Deferment e. In Forbearance f. Paid-in-Full g. Fully Discharged h. In Default **7. 2-Year Repayment Status for Parent PLUS Borrowers** a. Making Progress b. Not Making Progress c. Delinquent d. In Deferment e. In Forbearance f. Paid-in-Full g. Fully Discharged h. In Default **8. Average Outstanding Direct Loan Volume** **9. Average Outstanding Parent PLUS Loan Volume**
Census ACS	**1. Educational Attainment** a. Less than a High School Diploma b. High School Graduate c. Some College, No Degree d. Associate e. Bachelor f. Graduate **2. Median Earnings** **3. Median Earnings by Race** **4. Median Earnings by Educational Attainment and Sex** **5. Median Household Income**

	6. Percent of Workforce by Supersector a. Construction b. Education and Health Services c. Financial Activities d. Government e. Information f. Leisure and Hospitality g. Manufacturing h. Natural Resources and Mining i. Other Services j. Professional and Business Services k. Trade, Transportation, and Utilities **7. Percent of Workforce by Supersector by Sex** **8. Number of Workers by Supersector** **9. Civilian Labor Force Participation** **10. Civilian Unemployment Rate** **11. Unemployment Rate by Age** **12. Unemployment by Educational Attainment** **13. Home Ownership Percentage**
FSA Data Center	**1. Recipients of Undergraduate Unsubsidized Direct Loans** **2. Recipients of Undergraduate Subsidized Direct Loans** **3. Recipients of Graduate Unsubsidized Direct Loans** **4. Recipients of Pell Grants** **5. Recipients of Parent PLUS Loans** **6. Recipients of Grad PLUS Loans**

Variable Notes

Years:

All metrics are measured at the most recent year of data available for each data source, with a few notable exceptions:

- Repayment metrics from the College Scorecard are from 2019, due to the repayment pause/disruption caused by the COVID-19 pandemic.
- IPEDS only requires institutions to report age metrics in odd-numbered years. As a result, there is considerable selection bias in these variables in even-numbered years. For this reason, FTE enrollment by age is not included in the Trend Explorer. For even data reporting years (i.e., 2022), the previous odd year's (i.e., 2021) data are used for age metrics only.
- Because the Census does not collect the ACS at the same time as the decennial census, educational attainment data are not available for 2020 in the Trend Explorer.
- Because the ACS is not collected in U.S. territories, we use data from the territory census, which was last taken in 2020.

Timing of IPEDS Surveys:

In IPEDS, different survey components are collected at different points in the year. Survey components also represent different date ranges depending on the collection time and context of the survey being collected.[1] For the data included in our Explorer, we combine survey components based on the shared release year in the IPEDS Complete Data Files repository (2023 being the most recent complete repository). While 2023 data are provisional, we intend to update all data once final numbers have been calculated.

Institutional Sector:

- Instead of using the "sector" variable provided in IPEDS, we use the "preddeg" and "control" variables to create institutional sectors more aligned with the College Scorecard and other sources. We do this because IPEDS assigns "sector" based on the highest degree awarded rather than the most common degree awarded, resulting in community colleges with a single bachelor's degree program being categorized as a four-year institution.

Trend Explorer & Longitudinal Data:

- All cost measures available in the Trend Explorer have been adjusted for inflation to current dollars (for the most recent year in our data) using the Consumer Price Index (provided by the U.S. Bureau of Labor Statistics (BLS)). Net price measures that include income thresholds were not adjusted for inflation because the thresholds have remained unchanged throughout the decade.
- Measures without consistent longitudinal data (such as repayment measures) are not available in the trend explorer.

Full-Time Equivalent (FTE) Enrollment:

- FTE enrollment was calculated using the recommended estimates from the National Center for Education Statistics (NCES) glossary.

Tuition Discount Rates:

- Tuition discount rates were calculated by dividing total institutional aid by the total gross tuition revenue from the IPEDS finance survey. The variables used to calculate total gross tuition revenue were recommended in Cheslock (2019).[2] Total institutional aid was

[1] For a complete description of data collection and coverage in IPEDS, see their guide: https://nces.ed.gov/ipeds/use-the-data/timing-of-ipeds-data-collection.

[2] Cheslock, J. (2019). Examining instructional spending for accountability and consumer information purposes. *The Century Foundation*. Retrieved from https://tcf.org/content/report/examining-instructional-spending-accountability-consumer-information-purposes/.

calculated by summing restricted and unrestricted institutional aid, as demonstrated in Hillman (2012)[3] and Baum & Ma (2010).[4]

Enrollment by Age Group:

- We demarcate age groups as "24 & Under" and "25 & Up" based on the NCES definitions of "traditional" and "post-traditional" students.

Net Price:

- Net price information is available only for students in the Title IV financial aid cohort.

Graduation Rates at 150% Time:

- These rates represent only first-time/full-time degree-seeking students
- 150% time is calculated by using a six-year graduation rate for institutions in the four-year sector and a three-year graduation rate for institutions in the two-year sector

Census Educational Attainment:

- Educational attainment data from the Census are based on the ACS 1-year estimates.
- These data represent the share of attainment for the entire population (in a geography) aged 25 & up.

Census Workforce Variables:

- Workforce measures from the Census are drawn directly from published American Community Survey tables and reflect Census-defined populations, age ranges, and denominators. Different workforce variables may apply to different age groups or populations depending on how each concept is defined by the Census Bureau.
- Workforce measures are drawn from ACS 1-year estimates where available.
- Employment, unemployment, labor force participation, and related measures are reported for the civilian population only, consistent with standard Census definitions. Individuals on active military duty are excluded.
- Industry-of-employment measures are reported for employed workers and are presented both as counts and as shares of the workforce, depending on the underlying ACS table.
- Not all workforce variables are available with the same set of disaggregations. Some measures are published only for the total population, while others are available by select demographic or educational characteristics. The PNPI Explorer reflects the structure of the underlying ACS tables and does not create custom cross-tabulations beyond those provided by the Census Bureau.

---

[3] Hillman, N. W. (2012). Tuition discounting for revenue management. *Research in Higher Education 53*, 263-281. Retrieved from https://doi.org/10.1007/s11162-011-9233-4.

[4] Baum, S. & Ma, J. (2010). Tuition discounting: Institutional aid patterns at public and private colleges and universities, 2000-01 to 2008-09 (p. 8). New York: The College Board. Retrieved from https://research.collegeboard.org/media/pdf/trends-2010-tuition-discounting-institutional-aid-brief.pdf.

- Workforce variables apply to different population universes and age ranges depending on the underlying ACS table. The Explorer prioritizes disaggregations directly provided by the Census Bureau that can be applied uniformly at the national, state, territory, and congressional district levels.
- ACS uses the North American Industry Classification System (NAICS) to classify business establishments within its data. At their broadest level, they provide a set of 20 two-digit sector codes. Some government agencies, including the U.S. Bureau of Labor Statistics, have consolidated these sectors into ten "Supersectors," with an optional eleventh Supersector group for Public Administration. The PNPI Explorer uses the eleven-category Supersector classification, which is supported within ACS data.
- Educational attainment categories vary across workforce measures because the ACS publishes indicators in separate tables with different groupings. Some measures use the standard six-category attainment classification, while others combine or omit categories depending on how the metric is reported.
- Workforce variables use different age ranges based on Census definitions; labor force measures generally apply to ages 16 and older, while earnings and income measures are limited to older populations.
- Employment, unemployment, and labor force participation are reported as separate measures rather than a single employment status. This reflects Census definitions, as individuals who are not in the labor force are neither working nor actively seeking work and are therefore not counted as unemployed.
- Home ownership is included as a supplemental indicator of economic well-being. The American Community Survey provides limited direct measures of wealth, and home ownership is one of the few consistently available indicators across geographies and years.

Pell Grant Recipients:

- In IPEDS, Pell Grant Recipients are reported based on the Fall enrollment cohort. The FSA Data Center program volume files, however, provide an annual year-end summary report of the total number of Pell Grant Recipients. The FSA Data Center metric provides a more comprehensive view of Pell disbursements, so we have opted to calculate Pell Grant Recipients by dividing FSA Data Center recipient numbers by total undergraduate enrollment (as reported in the Student Financial Aid (SFA) IPEDS file), where possible. Figure 3 from Hillman (2018)[5] shows that this approach added nearly one million more recipients to the estimates. Because FSA Data Center volume files report information at the 6-digit OPEID, whereas IPEDS reports data at a mixture of 6- and 8-digit OPEIDs (depending on the specific way in which an institution reports), there are some cases where a campus may report more Pell recipients than undergraduates enrolled. This is because

[5] Hillman, N. W. (2018). Making the IPEDS Student Financial Aid Survey Data Meaningful. *National Postsecondary Education Cooperative.* Retrieved from https://nces.ed.gov/ipeds/pdf/NPEC/Data/NPEC_Paper_IPEDS_Studeto identify rows uniquelynt_Financial_Aid_2018.pdf.

Pell recipients are reported for all campuses in a single observation in the FSA Data Center, whereas enrollment is reported separately for each campus. In these few cases, we adjust the Pell recipient data in three ways:

- We ensure more coverage by using the larger number of recipients between the IPEDS and FSA Data Center recipient data; if either FSA Data Center or IPEDS is missing information for an institution, we use the other data source. We use IPEDS data if the FSA Data Center Pell recipient information exceeds the total amount of undergraduate enrollment. These adjustments prevent districts and sectors from reporting more than 100% Pell Grant recipients.

Aggregation Notes

Data from IPEDS, the Scorecard, and the FSA Data Center are at the institution level. While the FSA Data Center uses the OPEID to uniquely identify rows, both IPEDS and Scorecard use the IPEDS "UNITID" as their unique identifier (though the Scorecard provides a crosswalk between UNITID and OPEID from FSA Data Center data). Where possible, estimates were totaled at the geographic level before any calculations or aggregations were performed. For example, "Share of Black FTE Enrollment" is calculated by summing total Black FTE enrollment and total FTE enrollment at the geographic level, rather than averaging institution-level percentages. In some cases, aggregate measures were reported, along with the corresponding denominator (as with repayment metrics from the College Scorecard). In these cases, we disaggregated the measure, totaled these counts at the geographic level, and re-aggregated them. For instance, the share of undergraduates in the 2-year repayment cohort in default is reported for each institution in the College Scorecard, along with the total number of undergraduates in the 2-year repayment cohort. We multiply the share in default by the denominator, sum the total number in default at the geographic level, sum the denominator, and recalculate the share by geography.

When true aggregation was not possible, we used weighted estimates. Cost of attendance, tuition and fees, room, board, and other expenses were weighted by FTE enrollment. Net price measures were weighted using the net price cohort denominators provided in the College Scorecard. The average 3-Year Cohort Default Rate (CDR) was estimated using the CDR denominator provided by the College Scorecard. Median student debt (total by family income and completion status) could not be estimated beyond the institution-level estimates offered by the College Scorecard. For these metrics, we calculate a "median of medians." Since median estimates are resistant to outliers, we believe these "median of medians" are reliable and meaningful in our context.

Because of changes in IPEDS reporting, several institutions within the Pennsylvania State University system were manually assigned to their respective congressional districts.

Backend Methods

IPEDS:

Using Stata, we import the raw .csv files for each relevant IPEDS survey component and match them by UNITID. For files in long format (such as degrees conferred), data are reshaped to the UNITID level of analysis. This process is completed for each data year, after which the years are appended to a single Stata .dta file.

After combining the files, any variable calculations are made (such as generating tuition discount rates). We then keep only the necessary variables and remove all others (such as the web URL and the president's name). The Stata panel data file (all years) is stored in .dta and .csv formats.

College Scorecard:

Using Stata, we pull in the combined institution-level .csv file for the most recent cohort from the College Scorecard website. The College Scorecard already provides all its data at the UNITID level of analysis. We then import their historical data (provided as a zipped folder containing year-specific .csv files) and append the current year to all other years. After dropping unnecessary variables, we remove all “NA”, “NULL”, “PrivacySuppressed” (“PS”), and “PS” values from the retained variables and store them as missing values. By treating incomplete and privacy-suppressed data as missing, some measures may reflect small gaps in institutional reporting; however, these omissions have a minimal impact on overall geographic estimates. The Stata panel data file (all years) is stored in .dta and .csv formats.

Census ACS:

The Census Bureau provides an application programming interface (API) for each of its major data products, including the American Community Survey (ACS) 1-year, 3-year, and 5-year estimates. State, territory, and congressional district–level estimates are available through the 1-year and 5-year ACS products. Using Stata, we utilize the user-written packages jsonio and getcensus to access the Census API and retrieve a broad set of workforce-related measures for states, territories, congressional districts, and the nation. These measures include indicators of earnings and income, employment and labor force participation, unemployment, industry of employment, and related demographic breakdowns, along with the appropriate population denominators where required. Workforce variables are sourced from multiple ACS table families, each with its own population universe, age definition, and level of disaggregation. Availability of specific breakdowns is determined by the Census Bureau’s table structure rather than analytical choice.

In addition to core employment and earnings measures, select contextual indicators such as home ownership are included to provide additional insight into local economic conditions, given the limited availability of direct wealth measures in the ACS. Retrieved data are harmonized across years, used to calculate derived metrics and shares where applicable, and stored in a single longitudinal file spanning all available years. Final datasets are saved in both .dta and .csv formats at the geographic FIPS-level unit of analysis.

Federal Student Aid:

The FSA Data Center volume files on the FSA Data Center website are stored as Microsoft Excel files on a quarterly basis. In the fourth quarter of each year, the files provide an end-of-year annual summary for total program volume and recipients. These files report data at the OPEID level rather than the UNITID level. After parsing and cleaning the data for loan volume, grant aid, and campus-based programs for each year, the files are combined at the OPEID level, appended to create a panel (all years), and saved as .dta and .csv files.

Combined:

To combine these four cleaned panel files, we began with IPEDS, as it serves as our foundation for geographic identification and provides much of our data. We first merged Scorecard data, as that is the simplest crosswalk using UNITID and year as our two merging variables. A small number of institutions did not match, all of which were either not in the Title IV program participation universe or lacked matching congressional district, territory, or state geographic identifiers. These locations were dropped.

We next generated a string value of the OPEID in IPEDS/Scorecard to create a unique 8-digit number that can effectively match to FSA Data Center data. After merging the FSA Data Center panel using OPEID and year as our merge variables, a small number of institutions from FSA Data Center data did not match those in IPEDS. All of these institutions were small, for-profit campuses outside the analysis universe (based on our inclusion criteria) or were foreign campuses of larger institutions. These locations were dropped.

We next did a series of many-to-one merges using our institution-level file (IPEDS + Scorecard + FSA Data Center) to our Census ACS file; first, we merged on Congressional District ID and year, second, we merged on State FIPS and year, and third, we merged on a dummy indicator for national estimates and year. Because it was a many-to-one merge, institutions within the same congressional district shared the district's educational attainment values, institutions in the same state shared the state's values, and all institutions in the file shared the national values.

Finally, we trimmed and cleaned the combined data file (using our inclusion criteria). Once we had our cleaned, combined file, we split it into multiple .csv segments to reduce load time on our frontend. Each report type in the Explorer has a different underlying data file that contains only the necessary variables. These files have also been further aggregated. For all files except the Trend Explorer, the most recent year was retained. We then aggregated each variable at four geographic levels: national, state, territorial, and district. After aggregating, we split the file into geographic-specific forms, removed all duplicates (i.e., the national file contains just four rows, one for each institutional sector), and recombined the geographic files. This was completed for each of the nine report types (Topline, Enrollment & Access, College Cost, Completion & Attainment, Student Debt, Student Loan Repayment, Custom Report, Comparison Tool, and Trend Explorer). Each file was saved as some variation of explorer_topline.csv.

Frontend Methods

Using the .csv files generated by the backend method, we created multiple workbooks within a bespoke user interface to design and implement the tool's core functions. Each report type (Topline, Enrollment & Access, College Cost, Completion & Attainment, Student Debt, Student Loan Repayment, Custom Report, Comparison Tool, and Trend Explorer) was created as a separate data file and hosted on a custom server designed for PNPI.

Minority Serving Institutions

The PNPI Explorer provides a searchable list of institutions included in the user's selected geography. For each included institution, this list contains the institution's name, sector, and whether it meets the program area definition of various Minority Serving Institution (MSI) types. Eligibility was determined using the U.S. Department of Education's MSI Eligibility Matrix. Categories of MSIs included in the PNPI Explorer are: Asian American and Native American Pacific Islander Serving Institutions (AANAPISIs), Hispanic Serving Institutions (HSIs), Predominantly Black Institutions (PBIs), Historically Black Colleges and Universities (HBCUs), and Tribal Colleges and Universities (TCUs).

Acknowledgements

We would like to thank the members of our original technical review committee —Jinann Bitar, Dr. Nicholas Hillman, Nate Kelty, and Dr. Laura Perna — and members of our workforce technical review committee – Dr. Rajeev Darolia, Taylor Maag, Dr. Christopher M. Mullin, and Dr. Jared Colston – for their insights, feedback, and expertise.