Last Updated: 10/22/2024

The PNPI Explorer
Methodology Report

Purpose of the Explorer

The PNPI Explorer puts key data points directly in the hands of the policymaking community in an easily consumable form. Whether users are congressional staff, advocates, researchers, policy students, reporters, 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 a variety of federal sources, the PNPI Explorer allows users to create fully customizable higher education profiles at the national, state, and congressional district level—something that distinguishes our tool from others that are currently available. Users can examine how a given state, congressional district, or U.S. territory (geography) intersects with variables such as enrollment, race, degree conferred, and student debt; measure data against state and national averages; and see how variables have changed at the district, state, and national level over time. For all analyses, 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 using Stata and is an easily manageable workflow, allowing for minimal work to incorporate new data as they are released.

Institution Inclusion Criteria

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

- Institutions listed in FSA as being a campus outside of the United States;
- Institutions that did not match to 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);
- The institutions are in territories which are not exclusively affiliated with the United States (Marshall Islands and Federated Republic of Micronesia);
- Institutions that had 66%+ 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 Issues

Because most of our data come from the Integrated Postsecondary Education Data System (IPEDS), the College Scorecard, and the Federal Student Aid Data Center (FSA), the geography we are discussing is the congressional district or state in which that institution resides. When we discuss the share of students in the two-year repayment cohort in default, these are students who attended institutions in the given congressional district. That does not necessarily mean that those students still reside in the congressional district or state in question. While we limit the institutions in our data to those without large shares of online enrollment, there are also still online students included in these data. Therefore, the median student debt of an institution may include an individual not living in the same congressional district or state as the institution.

The second geographic issue lies with the nature of how IPEDS collects their data (because we match FSA and Scorecard contingent on IPEDS geographic identifiers). IPEDS requires institutions to report data based on the Program Participation Agreement (PPA); specifically, the reporting agent for the institution reports data as the campus where the PPA is held. There is considerable variation in the consequences of that requirement. Some institutions have multiple campuses that all report their data separately, thus they each have campus (and geography) specific data reported in IPEDS. Others report the data for all branches at the main campus, as in the case of Ivy Tech Community College in Indiana. Despite Ivy Tech having locations across the state of Indiana, IPEDS reports their data as all in Indianapolis, making geographic aggregation imprecise.

Most Recent Congressional Districts

Our data use the most recent congressional districts as the geographic basis of analysis. Though congressional districts are provided in IPEDS, the information is from older designations prior to the redistricting of the 2020 Census. To get and validate current congressional districts, we take the address provided in IPEDS for each institution and run it through Google’s Civic Information API which has addresses geocoded with representative and district information. We then cross-reference these new districts with the prior districts provided in IPEDS, highlighting any discrepancies. Among those locations where the new district is different for the most recent year, we randomly sampled addresses and manually checked the correct district using the Census Bureau’s My Congressional District look up feature. After confirming the accuracy of Google’s Civic Information API, we use the current congressional district for all locations.

For the Trend Explorer longitudinal data, we assign the current congressional district for all years after 2018 (the last year IPEDS provides an accurate congressional district).

Data Sources & Variables

Data come from four central public data sources: the Integrated Postsecondary Education Data System (IPEDS), the College Scorecard, the Census Bureau’s American Community Survey (ACS), and the Federal Student Aid Data Center (FSA). Below are the variable categories from each of the different sources. For a full list of the variable names, variable labels, and corresponding Tableau aggregate metric, please see 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 High School Diploma b. High School Graduate c. Some College, No Degree d. Associate e. Bachelor f. Graduate
FSA	**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 of the COVID-19 Pandemic.
- IPEDS does not require institutions to report age metrics except in odd-numbered years. Because of this, there is considerable selection bias in even-numbered years for these variables. For this reason, FTE enrollment by age is not included in the Trend Explorer. When this metric is reported, 2021 data are used in place of incomplete 2022 data.
- Because the Census does not collect the ACS at the same time as the decennial census, educational attainment data is not available for 2020 in the Trend Explorer.

Timing of IPEDS Surveys:

In IPEDS, different survey components are collected at different points of 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 (2022 being the most recent complete repository).

Institutional Sector:

- Instead of using the `sector` variable provided in IPEDS, we use the `preddeg` and `control` variables to create institutional sectors more in line with the College Scorecard and other sources. We do this because IPEDS assigns `sector` based on highest degree awarded rather than 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. Net price measures that include income thresholds were not adjusted for inflation because the income thresholds have remained the same for the entire decade.
- Measures that do not have consistent longitudinal data (such as repayment measures) are not available in the trend explorer.

[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.

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 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 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 the age groups as "24 & Under" and "25 & Up" based on the common definition of "traditional" and "post-traditional" students as seen in NCES reports.

Net Price:

- Net price information represents only students in the Title IV financial aid cohort.

Graduation Rates at 150% Time:

- These rates only represent 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

Pell Grant Recipients:

- In IPEDS, Pell Grant Recipients are reported based on the Fall enrollment cohort. The FSA program volume files, however, provide an annual year-end summary report of the total number of Pell Grant Recipients. The FSA metric shows a fuller picture of Pell disbursement, and so we have opted to calculate Pell Grant Recipients by dividing FSA

[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/.
[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.

recipient numbers by total undergraduate enrollment (as reported in the SFA IPEDS file) where possible. Figure 3 from Hillman (2018)[5] shows that doing this added nearly one million more recipients to the estimates. Because FSA 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 they are reporting Pell recipients for all campuses in a single observation in FSA, while enrollment is reported at 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 recipient data; if either FSA or IPEDS is missing information for an institution, we use the other data source; we use IPEDS data if FSA 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 FSA are at the institution level. While FSA 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). Where possible, estimates were totaled at the geographic level prior to any type of calculation or aggregation. For instance, "Share of Black FTE Enrollment" is not an averaged share among institution-level percentages; instead, total Black FTE enrollment and total FTE enrollment were summed at the given geography prior to calculating a percentage. In some cases, aggregate measures were reported but the denominator was also provided (as was the case with repayment metrics from the College Scorecard). In these cases, we disaggregated the measure, totalled these counts at the geographic level, and re-aggregated. For instance, the share of undergraduates in the 2-year repayment cohort in default is reported for each institution in the College Scorecard, as well as the denominator of total undergraduates in the 2-year repayment cohort for each institution. We multiply the share in default by the denominator, sum the total number in default at the geographic level, sum the denominator, and re-calculate the share by geography.

In cases where true aggregation was not possible, we used a number of weighted estimates. Cost of attendance, tuition & fees, room, board, and other fees 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) was not able to be estimated beyond the institution-level estimates provided by the College Scorecard. For these metrics, we calculate a "median of medians." Because median

[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_Student_Financial_Aid_2018.pdf.

estimates are not influenced by outliers in the distribution, we feel confident that these "median of median" estimates are useful and accurate for our context.

Other Notes

- In 2021-2022, University of Colorado Boulder did not report any State Appropriations, which is why their congressional district appears to have such low Appropriations per FTE.
- 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 pull in the raw .csv files for each relevant IPEDS survey component and match by UNITID. For files that are in long format (such as degrees conferred), data are reshaped to be UNITID-level of analysis. This process is completed for each 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 web URL and name of president). The Stata panel data file (all years) is then stored as .dta and .csv.

Scorecard:

Using Stata, we pull in the combined institution-level .csv file for the most recent cohort from the Scorecard website. Scorecard already provides all of their data at the UNITID-level of analysis. We then pull in their historical data (provided as a zipped folder containing year-specific .csv files) and append the current year with all other years. After dropping unnecessary variables, we remove all "NA", "NULL" and "PrivacySuppressed" or "PS" values from the kept variables and store them instead as missing. The Stata panel data file (all years) is then stored as .dta and .csv.

Census ACS:

The Census provides an API for each product, including the ACS for 1-year, 3-year, and 5-year estimates. State and congressional district level estimates are available in 1-year and 5-year estimates. Using Stata, we use the packages `-jsonio-` and `-getcensus-` to call the Census API and retrieve state, congressional district, and national estimates for the total number of people by educational attainment, as well as the educational attainment denominator. We then calculate the shares of educational attainment levels and store the single panel file (all years) as a .dta and .csv file at the geographic FIPS-level unit of analysis.

FSA:

The FSA volume files on their website are stored as Microsoft Excel 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 UNITID. 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 and appended to create a panel (all years) and saved as a .dta and .csv file.

Combined:

To combine these four cleaned panel files, we started with IPEDS, as that is our foundation of geographic identification and 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 did not have matching congressional district 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 match effectively to FSA. After merging with the FSA panel using OPEID and year as our merge variables, a small number of institutions from FSA did not match to IPEDS. All of these institutions were small for-profit campuses that were outside of 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) 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. There were no issues with missing values from this merge. Because it was a many-to-one merge, institutions within the same congressional district shared the same educational attainment values for the district, institutions in the same state shared the state 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 the file into multiple .csv segments in order to reduce the load time in our frontend (Tableau). 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 other than the Trend Explorer, the most recent year was kept. We then aggregated each variable at three geographic levels: national, state, 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 of the institutional sectors), and appended the geographic files back together. 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 each .csv produced in the backend method, we created multiple workbooks in Tableau to design and implement the core functions of the tool. 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 Tableau workbook and hosted on the Tableau Public server. After hosting the dashboards on the Public Server, the website itself was built in JavaScript, where we used the Tableau Embedding API to create the various functions available. The API allows us to create HTML and JavaScript form functions that speak directly to our embedded dashboards.

Acknowledgements

We would like to thank the members of our technical review committee- Jinann Bitar, Dr. Nicholas Hillman, Nate Kelly, and Dr. Laura Perna- for their insights, feedback, and expertise.