Data Quality Assessment – Integrated Postsecondary Education Data System (IPEDS) 2019-20 Final Release

This assessment focuses on two surveys in the database: 12-month Enrollment 2018-19 and Completions 2018-19.

Rubric for Assessing Dataset Quality

1. Relevance

- What is the total number of items relevant to credentials?
  There are around 150 relevant variables in the assessed surveys. Items most relevant to credential attainment include institution name, ID, location; 12-month enrollment number by gender, race/ethnicity, immigration status, and level of student; awards/degrees conferred by program type, award level, race/ethnicity, and gender; number of students receiving awards/degrees by award level, gender, age, and race/ethnicity.

- What are the measures of NDC attainment like?
  Certificates and degrees.

- Are there any indicators related to education attainment that are unique to this dataset?
  Enrollment, award/degree and student completion counts by program type, award level and demographic groups.

- Are there indicators of other phenomena that could be of sociological significance?
  Yes, there is rich information on demographics.

- What is the purpose of the dataset, and how closely does that purpose align with the following use cases? (Evaluate as relevant or not relevant.)
  IPEDS provides basic data needed to describe — and analyze trends in — postsecondary education in the United States, in terms of the numbers of students enrolled, staff employed, dollars expended, and degrees earned.
  a. Measuring the rate of attainment of credentials within the U.S. skilled technical workforce
     Relevant. IPEDS includes the number of degrees/awards conferred by program and the number of students receiving degrees/awards by institution, which can be further aggregated to measure the rate of attainment of credentials at higher levels.
  b. Measuring aggregate returns to credentials by credential type
     Somewhat relevant. IPEDS does not include data on labor market outcome but includes some information on students' subsequent enrollment in educational institutions.
  c. Identifying disparities by race and gender in the attainment of credentials
     Relevant. IPEDS tabulates completion measures by race/ethnicity and gender for each program/institution.
  d. Identifying which credentials are associated with the strongest labor market returns for individuals in the skilled technical workforce
     Not relevant. IPEDS does not include data on labor market outcome.
  e. Evaluating the effectiveness of public policies that support the attainment of credentials?

Relevant. IPEDS gathers information from every educational institution that participates in federal student aid programs. Its data supports policymaking regarding federal student aid and other policies targeting at postsecondary education attainment, especially those for specific demographic groups.

f. *Other examples we might add?*

2. Coverage

- What is the frame of reference for the dataset, what population does the dataset attempt to cover?
  IPEDS collects information from every college, university, and technical and vocational institution that participates in the Title IV federal student financial aid (FSA) programs. Institutions not eligible for federal student aid can request to be part of IPEDS.

- What is the number of cases, and how does that number compare to known estimates of the relevant population?
  The 2019-20 final release covers 6,559 institutions. This number is slightly more than the number of all Title IV institutions (approx. 6200) as some non-Title IV institutions also participate in IPEDS.

- How does the publisher of the data ensure that data is collected for cases that should be in the dataset?
  Since the completion of all IPEDS surveys is mandatory for all Title IV institutions, coverage for these institutions is almost 100 percent. Data collection procedures, including web collection and extensive email and telephone follow-up, are used to ensure high response rates. Since the implementation of the web collection in the 2000-01 cycle, Title IV institutional response rates for IPEDS surveys have ranged from 89 to over 99 percent. Imputation is performed to adjust for both partial and total nonresponse to a survey.

- Do cases that we believe should exist in the microdata actually exist in the data?
  Prior to the 2011 data collection, submission of new variables or surveys was optional for the first year of an institution's participation in IPEDS. Some data are only required in alternate years (e.g., enrollment by age in odd years), but schools may choose to submit every year.

- What percent of cases lack data for key variables of interest? (Direct assessment if not in metadata.)
  Most key variables have no missing value because nonresponse items have been imputed. Some variables are collected in odd/even years only and have high missing rates in non-collection years.

3. Granularity

- How granular (i.e., how many different categories exist, if not continuous) is data for key variables of interest (attainment, field of study, income)? What about for different levels of aggregation researchers might consider, such as geography, age, and race?
  - Institution location: Street address or post office box, city, state abbreviation, ZIP code, FIPS state code, Bureau of Economic Analysis (BEA) regions
  - Program type: 6-digit CIP code
  - Award level: Doctor's degree - research or scholarship/Doctor's degree - professional practice/Doctor's degree – other/Master's degree/Bachelor's degree/Associate

degree/Certificates of 2 but less than 4-years/Certificates of 1 but less than 2-years/Certificates of less than 1-year/Postbaccalaureate certificates/Post-master's certificates
    - Gender groups: Female/Male
    - Age groups: Under 18/18-24/25-39/40 and above/Unknown
    - Race groups: American Indian or Alaska Native/Asian/Black or African American/Native Hawaiian or other Pacific Islander/White/Two or more races/Unknown
    - Ethnicity groups: Hispanic/Non-Hispanic/Unknown
    - Immigration status: Nonresident alien/Others
    - Level of student: Undergraduate (including students in 4- or 5-year bachelor's degree programs, associate degree programs, vocational or technical programs below the baccalaureate, and students who have already earned a bachelor's degree but are taking undergraduate courses for credit) / Graduate (students who hold a bachelor's degree or above and is taking courses at the postbaccalaureate level, regardless of their enrollment status in graduate programs)

- Is this data granular enough (yes or no) to perform analyses for each of the use cases identified under “relevance”?
  Yes.

4. Timeliness

- How often is the dataset updated?
  Annually.

- Is data collected continuously or in waves? If in waves, what is the duration of those waves? Are some variables/records more frequently updated than others?
  In waves. There are 12 survey components in IPEDS and each is submitted annually in one of the three periods (Fall/Winter/Spring) in a collection year (cycle). Both 12-month Enrollment and Completions data are collected in Fall.

- What is the time lag between when an event occurs, when it is recorded, and when that data is available to researchers?
  In general, there is a one-year lag between event occurrence and data collection and a two-year lag between data collection and final release.12-month Enrollment and Completions data for the previous year are collected each Fall. Data collection starts in September and closes in November. Preliminary data is usually released 6 months after collection closes, and provisional data is released approximately 3 months after the preliminary release. Revised (final) data is released approximately 12 months after the provisional release. See here for details.

5. Integrity

- What are the risks to the integrity of this dataset?
  Data is reported by institutions or state agencies on behalf of the institutions, each of which could in theory have their own interests in the accuracy of data reported – especially if future federal funding may depend on the extent to which reported data demonstrates their performance.

- How are data outliers handled? (May be available from published documentation if not metadata.)

The web-based collection system automatically generates percentages for many data elements and totals for each survey page and compare current responses to previously reported data. Data elements are typically considered out of the expected range if the variance is greater than 25 percent. Survey respondents are allowed to correct errors detected by the system or to confirm that data is entered correctly or to key in a text message explaining why the data appear to be out of the expected range. Additionally, some outliers are marked "fatal" and must be corrected by the survey administrator rather than confirmed or explained by the respondent. Final quality control procedures are performed when all institutions have responded or data for them have been imputed.

- Were there changes to the dataset that may have resulted from political influence? If so, do those changes threaten the overall quality of the data?
  None that we are able to identify.

6. Accessibility

- How do researchers access this dataset?
  Complete or customized dataset of recent and past releases can be downloaded from the National Center for Education Statistics' website in CSV/Access/STATA/SAS/SPSS formats.
- Are any variables or cases withheld from researchers? If so, does that withholding or censoring affect researchers' ability to use the data?
  None that we are able to identify.
- Are there direct costs or indirect costs (e.g., training, resources) associated with accessing and using the data?
  Downloading and using the dataset is free.

7. Interoperability

- Is there a unique identifier for individual cases? If so, is it one that can be found in other datasets?
  Yes, each institution has a unique Unit ID. This ID is frequently used in other education statistics databases.
- Is it common? (e.g., a SSN might be higher value than an address, though even name/address might be sufficient to match in some cases).
  Yes.
- Do occupation and industry coding schemes correspond to commonly used frameworks such as O*Net and NAICS? If not, are they well documented in metadata?
  Yes, program CIP codes are available.

8. Suitability for Longitudinal Research?

- Is the metadata consistent over time, at least for key variables?

Metadata has been changed over time. New surveys and variables have been added, some variables are discontinued, and some variable definitions have changed.

- What is the construction of the dataset like? Is the microdata organized in "waves"? Can multiple observations for the same unit of analysis (i.e., person) over time be easily linked together?
  Each release includes multiple cross-sectional tables for different survey components. Historical releases are available and can be linked through Unit ID.

- How far back do administrative records from this dataset go?
  IPEDS data starts in 1986. Since 1993, IPEDS has surveyed the entire universe of postsecondary institutions. Prior to 1993, the coverage for private, for-profit, less-than-two-year institutions was about 15 percent. IPEDS replaced the Higher Education General Information Survey (HEGIS) in 1986. HEGIS collected data from 1966 to 1986 from a more limited universe of approximately 3,400 institutions. HEGIS data not on the IPEDS website are stored at the International Archive of Education Data, University of Michigan.