Data Quality Assessment – Noncredit Workforce Completers System (NWCS) of Maryland

Rubric for Assessing Dataset Quality

1. Relevance

- What is the total number of items relevant to credentials?
  There are 29 variables in the dataset, including data collection term and year; institution OPEID; student name, SSN/ITIN, local campus student ID, birthdate, ZIP code, birthdate, gender, race, ethnicity, Maryland residency status, citizenship status; course or course sequence name, type, CIP code, start and completion dates, award conferment date, instructional hours, and licensure/certification preparation status.

- What are the measures of NDC attainment like?
  Noncredit workforce training programs.

- Are there any indicators related to education attainment that are unique to this dataset?
  This dataset provides unique data on community college noncredit workforce training programs in Maryland.

- Are there indicators of other phenomena that could be of sociological significance?
  Yes, there is information on age, gender, race/ethnicity, Maryland residency status, and citizenship status.

- 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.)
  NWCS data is used for statewide reporting and analysis by Maryland state agencies on workforce outcome questions for completers of non-credit programs.
  a. Measuring the rate of attainment of credentials within the U.S. skilled technical workforce
     Relevant. Course completion and award conferment counts can be derived from the dataset and used to calculate relevant credential attainment rates in Maryland when combined with the state's population statistics.
  b. Measuring aggregate returns to credentials by credential type
     Somewhat relevant. NWCS itself does not include data on labor market outcomes but can be linked to wage data through SSN/ITIN.
  c. Identifying disparities by race and gender in the attainment of credentials
     Relevant. The dataset includes student race/ethnicity and gender.
  d. Identifying which credentials are associated with the strongest labor market returns for individuals in the skilled technical workforce
     Somewhat relevant. Researchers may use CIP codes to link occupation- or industry-level labor market data or use SSN/ITIN to link individual-level wage data.
  e. Evaluating the effectiveness of public policies that support the attainment of credentials?
     Relevant. NWCS data can be used to assess credential attainment among community college students and across different demographic groups, which supports policymaking regarding noncredit workforce training programs in Maryland.
  f. *Other examples we might add?*

Summary assessment: Based on the above, fill in a table describing relevance for each use case.

2. Coverage

- What is the frame of reference for the dataset, what population does the dataset attempt to cover?
  NWCS covers Maryland community college students who complete a noncredit workforce training course or sequence and are at least 16 years old at the beginning of course or sequence. Completers must have recorded grades for all courses. An eligible course or sequence is an approved noncredit certificate program leading to apprenticeships, employment, licensure, or job skill enhancement at a Maryland Community College.
- What is the number of cases, and how does that number compare to known estimates of the relevant population?
- How does the publisher of the data ensure that data is collected for cases that should be in the dataset?
  Institutional submissions are reviewed in two stages. The first stage performs Edit Checks. Edit checks confirm that only allowed values are entered for each data element. The second stage performs data validation and logic checks for data quality and data consistency.
- Do cases that we believe should exist in the microdata actually exist in the data?

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
  - Collection term: Fall/Winter/Spring/Summer/Cyber Warrior System/Annual MAPCS/Annual (academic year) used for DIS, ECS, FAIS and NWCS
  - Institution OPEID: 8-digit
  - ZIP code: 5-digit
  - Gender: Male/Female/Unknown, male assigned/Unknown, female assigned
  - Race: American Indian or Alaska Native/Asian/Black or African American/Native Hawaiian or other Pacific Islander/White/Multi-race/Unknown
  - Ethnicity: Hispanic/Non-Hispanic/Unknown
  - Citizenship: U.S. citizenship group consisting of U.S. citizens, U.S. nationals, resident aliens and other eligible non-citizens/Non-resident alien/Institution does not collect/Unknown
  - Residency status: Maryland resident/Non-Maryland resident
  - Course or sequence type: Business & Professional/Education/Health Care/Information Technology/Public Safety/Trades, Communications & Manufacturing/Transportation/Animal &

Plant Services/Culinary, Entertainment, Arts & Personal Services/Recreational and Fitness Professionals/Other/Not Applicable
    - Program CIP code: 4-digit
    - Licensure/certification preparation status: 1) prepares the student for licensure or industry certification through an exam or evaluation administered by third-party, 2) prepares the student for licensure or industry certification within the course through an examination, 3) prepares the student for licensure or industry certification within the course through general course content (no examination), or 4) the course/sequence does not lead to licensure and certification.

- Is this data granular enough (yes or no) to perform analyses for each of the use cases identified under "relevance"?
  Data granularity is generally high, but the CIP code is at 4-digit level only (maximum is 6-digit).

4. Timeliness

- How often is the dataset updated?
  Looks like 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. The collection year is from July 1 to June 30, and there are generally four collection terms (fall, winter, spring, and summer). Institutional data submission is due by December 1st.

- What is the time lag between when an event occurs, when it is recorded, and when that data is available to researchers?
  It could be over a year, depending on the timing of the event in question.

5. Integrity

- What are the risks to the integrity of this dataset?
  Data is reported by institutions, each of which could in theory have their own interests in the accuracy of data reported – especially if future 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.)
  We are not aware of any special handling of outliers.

- 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?

Researchers would have to apply by contacting the Maryland Higher Education Commission.

- Are any variables or cases withheld from researchers? If so, does that withholding or censoring affect researchers' ability to use the data?
  SSN/ITIN is scrambled to protect identity.
- Are there direct costs or indirect costs (e.g., training, resources) associated with accessing and using the data?
  No significant costs that we could identify.

7. Interoperability

- Is there a unique identifier for individual cases? If so, is it one that can be found in other datasets?
  Yes, student SSN/ITIN or local campus student ID is collected.
- 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?
  Yes.
- 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?
  Collection year and terms is provided for each record. SSN/ITIN and student ID can be used to link individual records over time.
- How far back do administrative records from this dataset go?
  The pilot collection started in FY2021 and will last through FY2022. The program is scheduled to be fully implemented in FY2023.