Data Quality Assessment – TrainingProviderResults.gov

Rubric for Assessing Dataset Quality

1. **Relevance**

- What is the total number of items relevant to non-degree credentials?
  51 variables in total, including training provider name, address, entity type; program name, description, location, URL, potential outcome type, associated credential, CIP code and title, O*NET codes, WIOA/non-WIOA tuition cost and supplies cost, program length in hours and weeks, prerequisites, format, number of all students/WIOA participants served/exited/completed/employed in the 2nd quarter after exit/employed in the 4th quarter after exit, number of all/WIOA exiters who attained a relevant credential within one year after exit, median earnings of all employed students in the 2nd quarter after exit, and the reporting state.
- What are the measures of NDC attainment like?
  Certificate, certification, license, or degree.
- Are there any indicators related to education attainment that are unique to this dataset?
  The potential outcome type and participant outcome variables for all students and WIOA participants of a program are unique to this dataset.
- Are there indicators of other phenomena that could be of sociological significance?
- 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.)
  The dataset is intended to:
  - help individuals make informed career training choices based on the program's completion and employment results;
  - help individuals make the best use of their Individual Training Account (ITA) funds;
  - assist American Job Center staff compare the quality of programs offered by approved training providers.

  a. Measuring the rate of attainment of non-degree credentials within the U.S. skilled technical workforce
     Relevant. Participation, exit, completion, and credential attainment counts are available for both all students and WIOA-participants of a program.
  b. Measuring aggregate returns to non-degree credentials by credential type
     Relevant. Employment and earnings statistics are available.
  c. Identifying disparities by race and gender in the attainment of non-degree credentials
     Not relevant. No race or gender information included.
  d. Identifying which non-degree credentials are associated with the strongest labor market returns for individuals in the skilled technical workforce
     Relevant. Employment and earnings statistics for multiple time periods after program exit are available.
  e. Evaluating the effectiveness of public policies that support the attainment of non-degree credentials?

Relevant. WIOA-specific statistics can be compared with total or non-WIOA statistics to evaluate the effectiveness of WIOA.

2. Coverage

- What is the frame of reference for the dataset, what population does the dataset attempt to cover?
  The dataset attempts to cover all job training programs eligible to be funded through WIOA.
- What is the number of cases, and how does that number compare to known estimates of the relevant population?
  As of May 22, 2022, the dataset covers 75,676 programs.
- How does the publisher of the data ensure that data is collected for cases that should be in the dataset?
  Data comes from annual state submissions. As required by WIOA, state reports should contain the performance information for all students and WIOA participants served by the program of study. DOL acknowledges that the policies and procedures for data collection and data quality assurances may vary from state to state and may be out of the Department's control.
- Do cases that we believe should exist in the microdata actually exist in the data?
  Not all training providers will be represented in the dataset for the following reasons:
  - The information reported by the provider may be suppressed to protect the personally identifiable information of training participants.
  - Providers that did not serve participants during the reporting period may not be included.
  - Newly added providers may not have data available at the time of reporting.
  - Some training providers, such as those providing certain work-based training, may not be required to report data.

  In addition, for the period from July 1, 2018 through June 30, 2021, for at least a portion of this period more than 30 states have received a waiver of the requirement to collect and report data on all students in a program. While states generally have reported the "all students" data they were able to collect, for some programs of study the "all students" data may be limited to WIOA students.
- What percent of cases lack data for key variables of interest? (Direct assessment if not in metadata.)
  Generally, missing rates are low for categorical variables but high for continuous variables. (For continuous variables, observations with value=-1 are counted as missing.)

Variable	Missing	%Missing	Variable	Missing	%Missing
training provider	0	0.0	associated credential	5,724	7.6
provider address	0	0.0	non wioa tuition cost	2,904	3.8
entity type	0	0.0	non wioa supplies cost	10,407	13.8
program name	2	0.0	cost per wioa	70,741	93.5
program description	19	0.0	program length hours	7,781	10.3
program url	23,964	31.7	program length weeks	6,466	8.5
address	0	0.0	program prerequisites	0	0.0
city	0	0.0	program format	0	0.0
state	0	0.0	cip code	0	0.0

zip	0	0.0	cip title	249	0.3
lat	0	0.0	onet 1	0	0.0
long	0	0.0	onet 2	0	0.0
program outcome type	0	0.0	onet 3	0	0.0
Variable	**Missing**	**%Missing**	**Variable**	**Missing**	**%Missing**
total served	55,086	72.8	wioa served	67,790	89.6
total exited	56,740	75.0	wioa exiters	69,570	91.9
total completed	58,763	77.7	wioa served with ita	70,618	93.3
completed percent	58,763	77.7	wioa exited with ita	71,721	94.8
total employed q2	60,481	79.9	wioa completed	70,551	93.2
total employed q4	61,309	81.0	wioa completed percent	70,551	93.2
emp percent q2	60,481	79.9	wioa employed q2	71,158	94.0
total emp perc comp q2	61,659	81.5	wioa employed q4	71,799	94.9
median earnings q2	61,907	81.8	wioa emp percent q2	71,158	94.0
total credential	64,887	85.7	wioa emp percent q4	71,799	94.9
			wioa credential	73,393	97.0
			wioa cred percent	73,393	97.0

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?
  - Entity type: Higher Ed: Associate's Degree/Higher Ed: Baccalaureate or Higher/Higher Ed: Certificate of Completion/National Apprenticeship/Private Non-Profit/Private For-Profit/Public/ Other.
  - Potential outcome type (combinations of the following): Industry-Recognized Certificate or Certification/Certificate of Completion of an Apprenticeship/License Recognized by the State Involved or the Federal Government/Associate's Degree/A program of study leading to a baccalaureate degree/IHE Certificate of Completion/Secondary School Diploma or Its Equivalent/Employment/Measurable Skill Gain Leading to a Credential/Measurable Skill Gain Leading to Employment.
  - Associated credential: type or specific name of the credential.
  - Prerequisite: None/High School Diploma or Equivalent/Associate's Degree/Bachelor's Degree/ Course(s)/Combination of Education and Course(s)
  - Format: In-person/Online, E-learning, or Distance Learning/Hybrid or Blended Program
  - CIP code: 6-digit.
  - O*NET code: 8-digit.
  - Location: latitude, longitude, address, city, state, 5-digit zip code.

- 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?
  Data comes from annual state submissions that occur each year by October 1 for state Eligible Training Provider Performance Reports (form ETA-9171).
- 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 (annually). Since data collection policies and procedures vary from state to state, records may be updated at different frequencies.
- What is the time lag between when an event occurs, when it is recorded, and when that data is available to researchers?
  All data in the downloadable public use file should be no more than one year old.

5. Integrity

- What are the risks to the integrity of this dataset?
  States agencies could theoretically 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 those agencies' performance.
- How are data outliers handled? (May be available from published documentation if not metadata.)
  Not clear. There are clearly some outliers in the dataset. For example, some programs require 67500 hours or 88920 weeks.
- 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?
  The dataset can be downloaded here in Excel format.
- Are any variables or cases withheld from researchers? If so, does that withholding or censoring affect researchers' ability to use the data?
  Data is suppressed from public view when any of the following occur:
  - Data submitted for the program contains sample sizes that are too small to protect Personally Identifiable Information;
  - No data were reported for the program; or
  - DOL identified significant data quality issues with the state submitted data.

  Such suppression may result in misrepresentation or selection bias that affect the validity of research findings.

- 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?
  No.
- 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).
  No UID, but linkage on individual cases is possible through provider and program names. While TPR provides no UID, some state reports have UID for providers and programs.
- 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?
  CIP and O*NET codes are included.

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?
  The dataset is cross-sectional. Theoretically, records for the same program over time can be linked together through program name, but historical datasets are not available on the TPR website.
- How far back do administrative records from this dataset go?
  WIOA [section 116(d)(4)] requires that state reports contain four years of data. As of the Program Year (PY) 2020 data submission (third year of reporting), state reports contain three years of data (PY 2018, PY 2019 and PY 2020).