Data Quality Assessment – WIOA Individual Performance Records (Public Use Data), PY2021 Q2

1. Relevance

- What is the total number of items relevant to non-degree credentials?
  The dataset contains a total of 289 variables. Items most relevant to NDC include an individual's , training program information, type and date of credential received, program entry and exit dates, employment status, employment industry and occupation, wage information, state/county/ZIP code, age at participation, gender, race/ethnicity, veteran status, disability status, other disadvantaged status, education level before participation, social welfare program participation information.

- What are the measures of NDC attainment like?
  A credential is classified as a diploma, degree, certification, certificate, or license.

- Are there any indicators related to education attainment that are unique to this dataset?
  Rich individual level data, including quarterly wage data before/after program participation.

- Are there indicators of other phenomena that could be of sociological significance?
  Yes, there is rich information on demographics, employment status, receipt of social welfare programs, and prior educational background.

- 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 purpose of this dataset is to assess the effectiveness of U.S. public investments in supporting unemployed and disadvantaged workers via the Workforce Innovation and Opportunity Act (WIOA).
  a. Measuring the rate of attainment of non-degree credentials within the U.S. skilled technical workforce
     Somewhat relevant. The dataset includes NDC attainment information but covers WIOA participants only.
  b. Measuring aggregate returns to non-degree credentials by credential type
     Relevant. The dataset includes rich wages information.
  c. Identifying disparities by race and gender in the attainment of non-degree credentials
     Relevant. The dataset includes race, ethnicity, and gender information.
  d. Identifying which non-degree credentials are associated with the strongest labor market returns for individuals in the skilled technical workforce
     Somewhat relevant. The dataset includes rich information on participant's wages before and after program participation.
  e. Evaluating the effectiveness of public policies that support the attainment of non-degree credentials?
     Relevant. WIOA is one of the most important public programs for supporting credential attainment and is likely to share characteristics with other existing and proposed public policy interventions.
  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?
  The WIOA Individual Performance Records cover all participants in WIOA-funded workforce development programs. The dataset includes most UI recipients in the United States (including DC and territories). Each quarterly file contains a rolling ten quarters of data. Q4 files contain data from Q1, Q2, and Q3 and are considered the annual dataset. Each file contains labor market outcome data for up to four quarters prior to quarter entry and four quarters after program completion.

- What is the number of cases, and how does that number compare to known estimates of the relevant population?
  The PY2021 Q2 release of the dataset contains 20,356,612 individual records. It contains information on individuals who were served by WIOA programs between July 1, 2019 and December 31, 2021. The number of records seems to be about right for the entire WIOA population relative to published estimates of the total number of individuals receiving WIOA support, considering that individuals are in the dataset for up to a year after program completion.

- How does the publisher of the data ensure that data is collected for cases that should be in the dataset?
  According to the Department of Labor, extensive training is provided for workforce professionals in state and local agencies to ensure that all program participants are recorded. We understand that DOL reviews data entries as they are received from individual states to assess accuracy.

- Do cases that we believe should exist in the microdata actually exist in the data?
  In Appendices B and C of the dataset, the data developer notes that certain states report no or very few individuals.

- What percent of cases lack data for key variables of interest? (Direct assessment if not in metadata.)
  Missing rate varies greatly by variable and state and is reported in the data appendices (available on the Employment and Training Administration's website).

Summary assessment: Qualitative description of evidence of completeness and/or steps taken to ensure completeness. Describe percent of variables of interest with missing data, any patterns we can infer as to the distribution of missing data. Evaluate whether the dataset is sufficient (yes or no) for each use case.

We see significant effort made to ensure completeness. Missing rate varies greatly by variable and state. Our analysis of the missing data did not find any specific patterns in the distribution of missing data, except that missing wage data tends to be more common for more recent quarters.

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?
  - Gender: Female/Male/Did not self-identify

- Race: American Indian or Alaska Native/Asian/Black or African American/Native Hawaiian or other Pacific Islander/White/Multiple-race selected/Did not self-identify
- Ethnicity: Hispanic/Non-Hispanic/Did not self-identify
- Education level: Secondary school diploma/Secondary school equivalency/Individualized Education Program/One of more years of postsecondary education/Postsecondary non-degree certification, license, or educational certificate/Associate degree/Bachelor's degree/Advanced degree/No educational level completed. Another variable records the highest school grade (0-12) completed at program entry.
- Veteran status: Yes/No/Not provided. Other variables include detailed information on veteran types.
- Disabled status: Yes/No/Not provided. Other variables include detailed information on disability type, type of service funds received, type of customized employment services received, work setting, and financial capability.
- Other disadvantaged status: Migrant and Seasonal Farmworker/TANF/SSI/SSD/SNAP/Pregnant or Parenting Youth/Foster Care Youth/Homeless/Ex-Offender/Low Income/English Language Learner/Basic Skills Deficient/Low Levels of Literacy/Cultural Barriers/Single Parent/Displaced
- Location: State, county, and ZIP.
- Program type: (Program leading to…) Industry certificate or certification/Registered apprenticeship certificate/License/Associate degree/Baccalaureate degree/Community college certificate of completion/Secondary school diploma or equivalency/Employment/Measurable skills gain
- Credential type: Secondary school diploma or equivalency/Associate degree/Baccalaureate degree/Occupational licensure/Occupational certificate/Occupational certification/Other recognized diploma, degree, or certificate/No recognized credential
- Industry and occupation codes: 6-digit NAICS, 6-digit CIP, 8-digit O*NET

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

Summary assessment: How do we rate the overall granularity of the data (high, medium, low)?

High.

4. Timeliness

- How often is the dataset updated?
  Quarterly.

- 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?
  Data is collected quarterly. Some variables are cumulative over a quarterly period, such as income (measured as the total of all earnings over three months), others are at a point in time at the end of the quarter (such as whether one is or is not employed and one's occupation). Exact dates are recorded for some events, such as starting or completing a credential.

- What is the time lag between when an event occurs, when it is recorded, and when that data is available to researchers?
  Events are recorded by state and local agencies on a quarterly basis. There is then a lag of one to two years between the time the data is reported to DOL and when it is made available to researchers.

Summary assessment: What is the length of the field period and the time between field and the availability of data to researchers?

The length of the field period is officially 90 days, but each data file contains data on events that occur before and after the field period. The lag in availability depends on a variety of factors but is usually between one and two years.

5. Integrity

- What are the risks to the integrity of this dataset?
  Data is handled by individual state agencies, 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 those agencies' performance.
- How are data outliers handled? (May be available from published documentation if not metadata.)
  In most cases, outliers are reported as-is. Very high incomes are top coded at $150,000 per quarter.
- 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.

Summary assessment: Describe any known risks to integrity we are able to determine from our research.

We do not identify significant risks to integrity.

6. Accessibility

- How do researchers access this dataset?
  Recent and past releases of the dataset in the CSV format and their appendices can be downloaded from the ETA website.
- Are any variables or cases withheld from researchers? If so, does that withholding or censoring affect researchers' ability to use the data?
  For confidentiality reasons, individual identifier is encrypted, and SSN is suppressed. The date of birth is suppressed and a calculated integer age is provided instead. Occupation and industry codes have been modified to display at a more general level or suppressed if there are fewer than 3 participants in a local area or an occupation/industry group in a local area. Wages have been rounded to the nearest whole dollar and randomly altered, but these adjusted wages retain the same underlying statistical properties in the aggregate as the actual wages. If a local area had 50 or

fewer exiters in a program year, those exiters are excluded from the file. This deletion does not apply to statewide programs.

ID encryption and SSN suppression affect the dataset's linkability with other data sources. Other data withholding or adjustment will not significantly affect researcher's ability to use the data.

- 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. Each quarterly release of public use data file contains over 20 million records and usually exceeds 8 GB. A computer/cloud computing service and a statistical software capable of reading and processing large files are needed.

Summary assessment: Is the data available to researchers? How do the hurdles to accessing data compare to other datasets we evaluate? Is the data access procedure consistent for all parts of the dataset, or are there pieces of the data that are more or less accessible?

The dataset is in a common format and free to download. Each data file is very large and requires sufficient computing capacity.

7. Interoperability

- Is there a unique identifier for individual cases? If so, is it one that can be found in other datasets?
  Yes, but it is unique to the dataset and encrypted in the public use data.
- 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.
- 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, O*NET, NAICS, and CIP codes are available.

Summary assessment: Are linkages possible on key variables or individual cases (yes or no)? Rate the potential for establishing meaningful data linkages for each use case (good, fair, poor).

Individual identifier is encrypted in the public use dataset so linkage with other datasets may not be possible. Linkage with industry- or occupation-level data is straightforward with NAICS, CIP and O*NET codes available.

8. Suitability for Longitudinal Research?

- Is the metadata consistent over time, at least for key variables?
  Yes. Quarterly data files on the ETA website are revised to reflect the newest Participant Individual Record Layout (PIRL), the major system used for WIOA individual record reporting. PIRL reporting began in 2016 and was modified in 2018, 2020, and 2021.

- 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 published in a separate file for each quarter. As the individual identifier is encrypted and SSN is suppressed in the dataset, it is not possible to link data from multiple quarters.

- How far back do administrative records from this dataset go?
  The oldest release available is PY2017 Q4, covering individuals served by WIOA programs between July 1, 2016, and June 30, 2018.

Summary assessment: Identify the length of time covered by the dataset (and the consistency of data collection over time) and rate as shorter or longer than other datasets. Objectively assess fit between time covered by data and time period of interest for each use case.

Combining all currently available data releases, the dataset covers individuals served by WIOA programs between July 1, 2016 and December 31, 2021.