National Center for Science and Engineering Statistics

Title: Opportunities to Understand the Skilled Technical Workforce (STW) Through Improved Administrative Datasets

Date: November 2023
Final Report

Contractor Awardee: George Washington Institute of Public Policy and the Center for Regional Economic Competitiveness
Contract Number: 49100421C0020

OPPORTUNITIES TO UNDERSTAND THE SKILLED TECHNICAL WORKFORCE (STW) THROUGH IMPROVED ADMINISTRATIVE DATASETS

Final Working Paper

NOVEMBER 2023

This report was produced by the George Washington Institute of Public Policy and the Center for Regional Economic Competitiveness for NCSES
BAA #49100421C0020

Co-Authors: Allison Forbes, Kyle Albert, Andrew Reamer
Research Team: Nichelle Williams, Ziyuan Wang, Anika Rahman, Valrie Eisele, Yuchen Xu

Table of Contents

PROJECT BACKGROUND

The National Center for Science and Engineering Statistics (NCSES), George Washington University's Institute for Public Policy (GWIPP), and the Center for Regional Economic Competitiveness (CREC) partnered to map how administrative data can contribute to statistics on the Skilled Technical Workforce (STW)and Non-Degree Credentials (NDCs). GWIPP and CREC pilot tested a data quality assessment approach which aims to increase the confidence of researchers in utilizing administrative data to study the STW. The GWIPP-CREC team assessed 20 potential datasets, created a metadata repository, and catalogued 350 relevant variables across 15 of those datasets. Twelve administrative datasets were included in a more detailed data quality assessment and six were used to identify the opportunities described in this paper for inter-agency collaboration.

This paper is the second in a series of three papers submitted to NCSES in 2023. Companion papers address "Potential Uses of Administrative Datasets to Complement and Inform the National Training, Education, and Workforce Survey" and "An Exploration of Re-Employment Using the Participant Individual Record Layout."

OUR APPROACH

In this paper, we ask two research questions of interest to the National Center for Science and Engineering Statistics (NCSES) in the U.S. National Science Foundation (NSF) and the broader research community: What is the quality of administrative datasets to describe the Skilled Technical Workforce (STW)[1] and how might the datasets be improved for statistical purposes? This exploration is timely as policymakers are considering alternatives to survey statistics identifying labor market trends affecting this critical segment of the labor market.

FINDINGS

The NCSES can cohere decentralized data collection processes by setting an inter-agency agenda to describe the STW. Inherent characteristics of administrative data make it difficult for scholars and statisticians to conduct research and produce statistics. A federal research strategy to identify and support the STW could advance immediate opportunities for interagency collaboration and begin to address outstanding challenges, including:

[1] The NSB (National Science Board) report titled "The Skilled Technical Workforce: Crafting America's Science & Engineering Enterprise" (2019) defines the Skilled Technical Workforce (STW) as "individuals who utilize science and engineering skills in their jobs but do not have a bachelor's degree." See *The Skilled technical workforce: Crafting America's science & engineering enterprise* (NSB-2019-23), National Science Board, (2019), https://nsf-gov-resources.nsf.gov/nsb/publications/2019/nsb201923.pdf.

- There are immediate opportunities for federal agencies to utilize instructional and occupational codes to identify STW-relevant training and describe STW participation in major public education and workforce programs.
- To remain nimble and prepare to address public policy priorities, federal agency leaders can begin to align and improve data collection for key units of analysis (participants, credentials, training programs, training providers) across various data collection efforts.
- With an inter-agency working group established, federal policy priorities can galvanize efforts to address persistent data collection and interpretation issues and improve information available for key units of analysis.
- An inter-agency effort could engage scholarly and applied research partners, providing microdata access for the purpose of describing STW career pathways and barriers to expanding the STW.

IDENTIFYING THE SKILLED TECHNICAL WORKFORCE, RELEVANT TRAINING AND CREDENTIALS

Over the past decade, the visibility of the Skilled Technical Workforce (STW) in U.S. industries and innovation systems has increased.[2,3] The National Science Board (NSB) recognizes that almost 17 million technical workers without a bachelor's degree are uniquely positioned to ensure the integration of new industrial and technological processes across their various fields and sectors, from healthcare to manufacturing.[4] These workers are critical to staffing tech-savvy businesses and deploying critical technologies.

Policymakers have access to important but limited information on the training and credentialing of the STW. Federal statistical surveys provide information on the total number and demographics of skilled technical workers, defined by occupation and education in the American Community Survey (ACS) or Current Population Survey (CPS). Only three questions in

[2] The National Science Board defines the Skilled Technical Workforce (STW) as "individuals who utilize science and engineering skills in their jobs but do not have a bachelor's degree". See *The Skilled technical workforce: Crafting America's science & engineering enterprise* (NSB-2019-23), National Science Board, (2019), https://www.nsf.gov/nsb/publications/2019/nsb201923.pdf.

[3] Federal agencies and national organizations are bringing new programs online to address the need for talent. Examples include the workforce requirements attached to incentives under the CHIPS Act for semiconductor manufacturing; the U.S. Department of Commerce's new SelectTalentUSA program, which seeks workforce solutions for foreign companies locating operations in the U.S.; and a recent White House report on the need for public health workers. See *U. S. Departments of commerce, labor, and education announce SelectTalentUSA, new partnership to increase quality jobs through FDI*, U.S. Department of Commerce, (2023), https://www.commerce.gov/news/press-releases/2023/05/us-departments-commerce-labor-and-education-announce-selecttalentusa; "U.S. Departments of commerce, labor, and education announce SelectTalentUSA, new partnership to increase quality jobs through FDI", U.S. Department of Commerce, (2023), https://www.whitehouse.gov/wp-content/uploads/2023/04/PCAST_Public-Health-Report_May2023.pdf.

[4] "Science and Engineering Labor Force," *Science & Engineering Indicators*, National Science Board, (2020). The definition for the STW used in the Indicators report builds on the definition introduced in Jonathan Rothwell's 2015 publication "Defining Skilled Technical Work." The National Center for Science and Engineering Statistics (NCSES) has added to Rothwell's definition by adding occupations that NCSES designates as S&E and S&E-related.

the CPS address non-degree credentials (NDCs).[5] **This will change with data from the new National Training, Education, and Workforce Survey (NTEWS),**[6] **designed to better capture STW career pathways,** including STW employment, education, credentials, and work experiences.

This data is important to fortify STW training and credential pathways and prepare a workforce capable of integrating new technologies in the workplace. A stronger STW is critical to national competitiveness, security, and the research enterprise. In 2019, NSB recommended action to: a) leverage the portfolio of federal investments, b) build partnerships across educational institutions that tailor local STW programs to local needs, and c) raise awareness of STW career and educational pathways to address misperceptions that a four-year degree is the only way to a rewarding career. Each of these actions requires data to guide decisions and demonstrate progress.

With the launch of the NTEWS, the new frontier is leveraging administrative data. Administrative data is collected for non-statistical purposes and can be found across government and the private sector. Typically, administrative data is collected or generated while performing administrative tasks, such as licensing vehicles, treating patients, conferring degrees, collecting taxes, regulating program compliance, or assessing program efficacy. Federal agencies have various initiatives underway to expand and enhance administrative data collection and analysis.[7,8,9] We expect that federal, state, and extra-governmental initiatives will make more

[5] From 2010 through 2018, the federal government made tremendous strides in collecting data on non-degree attainment through the Interagency Working Group on Expanded Measures of Enrollment and Attainment (GEMEnA), which facilitated the placement of indicators of non-degree attainment on a range of federal surveys. While these surveys are costly to administer and have inherent limitations in their sampling frame and the level of detail they can collect on each respondent and credential, they allow for nationally representative descriptions of the STW. See, "*The Skilled technical workforce: Crafting America's Science & Engineering Enterprise* (NSB-2019-23)", National Science Board, (2019), https://www.nsf.gov/nsb/publications/2019/nsb201923.pdf and; "*The STEM labor force of today: Scientists, engineers, and skilled technical workers*", National Science Foundation, (n.d.), https://ncses.nsf.gov/pubs/nsb20212/u-s-stem-workforce-definition-size-and-growth.

[6] This survey was in the field April 2022 to October 2022. The results for the 2022 survey cycle were released in 2024 and are made available to the public through the NTEWS web page. Update: NCSES released initial data tables from NTEWS in January 2025: https://ncses.nsf.gov/surveys/national-training-education-workforce/2022#data. Analysis of the data was released in Fall 2025, see: New Pilot Data on the Prevalence of Work-Related Credentials among STEM Workers from the National Training, Education, and Workforce Survey, National Science Foundation, https://ncses.nsf.gov/pubs/nsf25352.

[7] A project of National Academy of Sciences, Engineering, and Medicine. See: *Toward a vision for a new data infrastructure for federal statistics and social and economic research in the 21st century*,(n.d.), https://www.nationalacademies.org/our-work/toward-a-vision-for-a-new-data-infrastructure-for-federal-statistics-and-social-and-economic-research-in-the-21st-century.

[8] ITIF discusses federal data strategy announced 2019. See, Egan, E., *Reviving and reimagining the federal data strategy for mission success.* Information Technology & Innovation Foundation, (2023), https://www2.itif.org/2023-federal-data-strategy.pdf.

[9] *Chief data officer council: Progress in strengthening federal evidence-based policymaking*, U.S. Government Accountability, (n.d.), https://www.gao.gov/products/gao-23-105514.

data available in upcoming years.[10,11,12] These innovations in national information systems can help to identify effective STW education, training, and employment programs.

THE POTENTIAL FOR ADMINISTRATIVE DATA TO DESCRIBE THE STW

Administrative data enables strategic decision making at Federal, state, and local levels of government, within Territorial and Tribal Nations, and across the private sector. Administrative data can unlock valuable knowledge and drive modernization across public, non-profit, and private sectors. To further leverage administrative data for public policymaking, improved public data collection and analysis processes are needed.

Innovations in the collection and availability of high-quality administrative data are increasingly important across all levels of government, in part because of declining response rates in federal surveys. The Foundations for Evidence-Based Policymaking Act of 2018 (Evidence Act) encourages federal statistical agencies to explore data-leveraging opportunities through administrative and other supplemental data. The Act recognizes that administrative data may have more complete population coverage, lower data collection costs, reduced respondent burden, and can lead to better data quality. In accordance with the Evidence Act, agencies publish evaluation plans and establish advisory committees on data for evidence building to review, analyze, and make recommendations on how to promote the use of Federal data for evidence building.[13]

---

[10] For example, the **Coleridge Initiative's Administrative Data Research Facility (ADRF)**, which supports data sharing and analysis between state agencies, was established under guidance from the Census Bureau with funding from the Office of Management and Budget to inform the decision-making of the Commission on Evidence-Based Policy. Despite its currently limited scope for eligible participation and state coverage, this initiative demonstrates the potential for administrative data to inform policymaking at the state level - https://coleridgeinitiative.org/administrative-data-research-facility. See also a May 2023 update on the **U.S. Chamber of Commerce Foundation JEDx** initiative https://www.uschamberfoundation.org/workforce/the-potential-of-jedx-to-reduce-employer-burden-by-consolidation-reporting-to-government-initial-observations; a February 2022 report from the National Skills Coalition on relevant innovations in state-level data systems https://nationalskillscoalition.org/wp-content/uploads/2022/03/Emerging-Innovations-Final.pdf; and the **Workforce Almanac Data Portal** (https://workforcealmanac.com/) from the Harvard Project on Workforce, an open-source directory that visualizes data on nearly 17,000 workforce training providers across the United States, including provider names, location, and types.

[11] To address this issue, the Council of State Government (CSG) provides a Data Standard for individual-level data capture. To ease the process of data collection amongst multiple governmental partners, CSG has worked with Federal agencies to establish a standard naming convention and data fields and with states and localities to improve data quality and process efficiency. This work to align and connect could and should go further.

[12] Many new, relevant federal programs have been created and many are now more visibly connected to developing and utilizing the critical skills of the STW. Find a catalogue of federal programs from May 2023 here https://bpb-us-e1.wpmucdn.com/blogs.gwu.edu/dist/b/3597/files/2023/06/Skilled-Technical-Workforce-Programs.pdf - with updates accessible via the George Washington Institute for Public Policymaking (GWIPP) website https://gwipp.gwu.edu/administrative-data-repository-skilled-technical-workforce.

[13] Public Law, 115-435, 132 STAT.5529.https://www.congress.gov/115/plaws/publ435/PLAW-115publ435.pdf.

Federal agencies are aware of administrative data applications and issues, and existing expertise can be leveraged to understand the STW. For example:

- The U.S. Department of Labor (DOL) Chief Evaluation Office Administrative Data Research and Analysis portfolio of projects leverages administrative datasets held by DOL and other federal agencies for policy analysis. DOL uses its own administrative data to conduct sector-specific research in addition to using interagency administrative data collections such as the National Directory of New Hires. For example, one study examined job training trends using survey and administrative data.[14]
- The U.S. Census Bureau's Center for Administrative Records Research and Applications (CARRA) published research leveraging "the Data Linkage Infrastructure, the Census Bureau's clearing house for administrative, census and survey data" from 2014 to 2021. While all 60 of these working papers examined the effectiveness of administrative datasets and three evaluated administrative data for the National Survey of College Graduates (NSCG), another NCSES workforce survey,[15] none of them addressed the STW.
- The U.S. Department of Education's National Center for Education Statistics (NCES) collaborated with researchers and statisticians in several ways to examine and improve its own administrative dataset, the Integrated Postsecondary Education Data System (IPEDS), via the National Postsecondary Education Cooperative, the IPEDS Technical Review Panel, the NCES Data Institute, and through projects with individual institutional research professionals.[16] Statistics from IPEDS are regularly included in NCES reports.[17]
- The DOL Enterprise Data Strategy focuses on more consistent and effective data governance[18] and the 2023 U.S. Department of Education Data Strategy, as part of the agency's data access goals, aims to expand secure access to personally identifiable

[14] The report describes that wage record access is an issue across five databases and federal and military employment exclusion from wage records is an issue as well. They describe issues with reliability in self-employment reports to IRS that are integrated. Various datasets reviewed update their data at different points in time. See, Mastri, A., Rotz, D., & Hanno, E. S. (2018). *Comparing job training impact estimates using survey and administrative data*. Mathematica Policy Research. https://www.dol.gov/sites/dolgov/files/OASP/legacy/files/WIA-comparing-impacts.pdf.

[15] *Center for Administrative Records Research and Applications (CARRA) working papers,* Census.Gov, U.S. Census Bureau, https://www.census.gov/library/working-papers/series/carra-wp.html. Evaluating Administrative Records to Inform Measurement Error Properties of National Survey of College Graduates Estimates: Employment History and Firm Characteristics (2021), Evaluating Administrative Records to Inform Measurement Error Properties of National Survey of College Graduates Estimates: An Analysis of the NSCG-LEHD Earnings Ratio (2021), and Evaluating Administrative Records as a Potential Sample Frame for the National Survey of College Graduates (2021).

[16] Learn more about partnerships to explore and improve the Integrated Postsecondary Education Data System https://nces.ed.gov/ipeds/join-in.

[17] IPEDS has three collection periods annually, which are during Fall, Winter and Spring. NCES releases reports based on data collected through IPEDS throughout the year with several variations. Some reports are released annually, while some others are released on a less frequent basis. For example, the Digest of Education Statistics publishes information from pre-k through graduate school regarding education attainment, labor force status by educational attainment, occupations and earnings by educational attainment etc https://nces.ed.gov/programs/digest/.

[18] DOL Data Strategy. See, *Data strategy*, U.S. Department of Labor Office of Data Governance, (n.d.), https://www.dol.gov/agencies/odg/strategy.

information (PII) for external stakeholders, noting the relevance of student financial aid data for research purposes.[19]

The NCSES at NSF data strategy explicitly mentions using survey and administrative data in concert to better understand the training and credentialing of the STW. Other federal agencies have laid the groundwork to make important contributions to this effort. **Federal statistical agencies have experience with survey and administrative data and could apply similar or new approaches to identify and measure the STW**.[20]

Due to previous efforts by multiple federal agencies to collaborate and combine information, we have examples of high quality and useful administrative datasets that are available to the public. The U.S. Census' Longitudinal Employer-Household Dynamics (LEHD) exemplifies the potential for administrative data to inform public policymakers by capturing detailed information about local economies with the expressed goal of providing indicators needed by state and local authorities. LEHD (see fig. 1) is the result of an intergovernmental partnership launched in 1999. The Unemployment Insurance (UI) employment and earnings data from states, upon which the program relies, provides nationwide coverage and is integrated into a variety of federal data products (e.g., LEHD) and performance reporting processes that require employment information.

Given that the current UI reporting system for employers does not uniformly identify the job titles or occupations of employees, LEHD does not currently enable data products with nationwide information on the employment and training of the STW. Most states' Unemployment Insurance (UI) employment and earnings data, which underpins the Census LEHD program, includes the industry of the employer but not the job title of the employee, making it difficult to identify the STW. Long-range efforts are needed, and are underway, to "enhance" wage record data collection with additional information.[21] In the meantime, other datasets with information on job title or occupation will need to be linked to employment data to complete the picture.

The LEHD Post-Secondary Employment Outcomes (PSEO) program[22] provides a model for generating useful statistics by pairing UI wage records with information on educational programs, with its focus on linking employment outcomes to the completion of degree and certificate programs, including those relevant to the STW. PSEO contains experimental data from educational institutions in a growing number of states. With states' permission, Census matches educational information on program completers with wage records to produce

---

[19] *Data Strategy,* U.S. Department of Education, (2023), https://www.ed.gov/media/document/us-department-of-education-data-strategy.

[20] The most recently published evaluation plan (FY24) for the Department of Labor includes research on credential attainment through a 4-year study under ETA's Job Corps program and explores strategies for credential-focused curriculum through a 5-year evaluation of ETA's Strengthening Community College (SCC) grant program. See *U.S. Department of Labor fiscal year 23-24 evaluation plan,* (n.d.), https://www.dol.gov/sites/dolgov/files/evidence/DOL-CEO-FY-2023-2024-Evaluation-Plan.pdf.

[21] See the January 2022 state-by-state inventory for the BLS Labor Market Information Oversight Council and LMI Institute. *An Inventory of Employee-Specific Data Collected on Unemployment Insurance Wage Records, (2022)*. https://www.bls.gov/advisory/bloc/ui-wage-records-report-january-2022.pdf.

[22] U.S. Census Bureau Center for Economic Studies. (n.d.). https://lehd.ces.census.gov/data/pseo_experimental.html

education and employment statistics at the program level – again, linking to industries of employment not occupations or job titles, given the limitations of wage record data. However, the Classification of Instructional Programs (CIP) codes can be used to identify educational programs relevant to the STW. CIP codes can be linked to Standard Occupational Codes (SOC) to help identify programs relevant to STW employment.

Figure 1. The U.S. Census Bureau LEHD Program

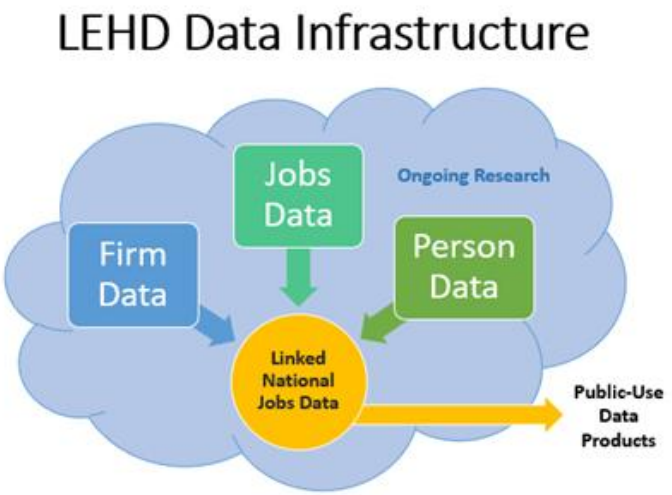


Image from U.S. Census Bureau - https://lehd.ces.census.gov/

> Under the Local Employment Dynamics (LED) Partnership, states agree to share Unemployment Insurance earnings data and the Quarterly Census of Employment and Wages (QCEW) data with the Census Bureau. **The LEHD program** combines these administrative data, additional administrative data and data from censuses and surveys. From these data, the program creates statistics on employment, earnings, and job flows at detailed levels of geography and industry and for different demographic groups. In addition, the LEHD program uses these data to create partially synthetic data on workers' residential patterns.[23]

The U.S. Department of Labor's Employment and Training Administration (ETA) maintains several additional administrative datasets that contain information relevant to the training and credentialing of the STW, including two datasets on program participation that can be published with deidentified individual records. In 2023, a data dashboard published by the Office of Apprenticeship (OA) Registered Apprenticeship Partners Information Database System (RAPIDS) provided a view to previously unpublished participant demographics from 2014 to 2023, and a data dashboard for the Workforce Innovation and Opportunity Act (WIOA) Participant Individual Record Layout (PIRL) made 2017-2021 data more visible to the public. [24] Both these datasets are used to generate statistics useful for performance evaluation. As with the PSEO, states play an important role in producing these data products.

---

[23] *Longitudinal employer-household dynamics*, U.S. Census Bureau Center for Economic Studies, (n.d.), https://lehd.ces.census.gov/.
[24] See dashboards at https://www.apprenticeship.gov/data-and-statistics.

Underlying datasets for both RAPIDS and PIRL include individual records that could be linked to other information on program participation at the state or national level to better understand the characteristics of workers in STW occupations, their training and education. PIRL is linked to wage records for employment outcome information and RAPIDS could be similarly linked to create more robust information on employment results.

These and other federal administrative datasets are used to document education and training programs nationwide and could be applied to understand the STW and STW-relevant training. Administrative datasets from ETA can be used to describe non-degree credentials (NDCs) and relevant training programs – as demonstrated in a series of "Counting Credentials" reports by Credential Engine (2017-2022) and in creating the Workforce Almanac by The Project on Workforce at Harvard University. Both efforts make use of the Eligible Training Provider Performance Results (ETPPR), RAPIDS, and IPEDS datasets.[25,26] The Workforce Almanac also uses the IRS Exempt Organizations Business Master File to find additional non-profit educational institutions. The Counting Credentials reports match and deduplicate the IPEDS and ETPPR lists for the purpose of counting unique certificate programs nationwide, and a matched dataset is available via the ETA CareerOneStop Local Training Finder. The Workforce Almanac project matches and deduplicates across all four of their data sources to provide a searchable list of training providers nationwide. Many of the NDCs counted in the Counting Credentials report can be linked to information on occupation, which could help to identify the STW.[27]

In the next section, we describe how a data quality framework can be applied across administrative datasets, including several mentioned above, to identify their fit for describing STW training and credentialing and generating useful statistics for a variety of use cases.

DATA QUALITY ASSESSMENT ACROSS STW NDC ADMINISTRATIVE DATASETS

It is important to be realistic about what is possible with respect to assessing data quality. In the context of applying a total error framework to non-survey data, Amaya et. al. (2020: 116) write:

> *"The information required to conduct the ideal investigation is rarely, if ever, available. In some cases, this is because the truth may be un-knowable, the information is proprietary, or the effort required would be cost prohibitive. Instead, we must use what is available and attack the problem from several different angles, relying on the resulting big picture as opposed to the individual*

[25] The Project on Workforce, "Methodology," The Workforce Almanac, accessed November 22, 2023, https://workforcealmanac.com/methodology.
[26] Counting U.S. Postsecondary and Secondary Credentials, Credential Engine, (2022). https://credentialengine.org/resources/counting-u-s-secondary-and-postsecondary-credentials-report/.
[27] Ibid., see Appendix C, 60-61.

Administrative data is collected for programmatic purposes (e.g., tracking program participants and outcomes) rather than for statistical or research purposes. Most research on the quality of administrative datasets to date has occurred in the context of medicine and public health, an environment relatively rich in records (e.g., medical charts, billing records) that are largely digitized, but often in "walled" systems such as individual hospitals or state healthcare agencies. Efforts to develop rubrics to evaluate the quality of administrative datasets have been led by national statistical agencies in the U.S. and abroad.[28]

Various U.S. agencies have published standards for data quality and data acquisition. Some include methods to improve administrative datasets.

- The U.S. Census Bureau published a report in November 2022, articulating specific guidelines regarding assessing the quality of administrative data in a Census cycle.[29] The report outlined specific tools, indicators, and dimensions that agencies need to be mindful of in the evaluation process. This report was a positive sign that agencies creating these datasets are taking actions to improve the data quality and that a toolkit exists.
- The Federal Committee on Statistical Methodology (FCSM) published a framework for data quality assessment in September 2020, defining data quality using three domains: utility, objectivity, and integrity. The report further developed eleven dimensions within three domains to illustrate specific aspects of data quality to be considered.[30]
- The American Community Survey Office published a report in November 2018 establishing 12 guiding principles in assessing appropriate administrative dataset sources for ACS usage.[31] Among those 12 guiding principles, five core principles are coverage, quality, conceptual alignment, temporal alignment and impacts on estimates.
- In 2017, the Administrative Data and Research Analysis (ADRA) program of the Department of Labor contracted with the Mathematica Policy Research to publish a report discussing key determinants for administrative data sources on earnings, in which data access, coverage, reliability, and periodicity were four core determinants.[32]

---

[28] Researchers affiliated with the statistical and/or census agencies of several countries – including the United States, Sweden, the Netherlands, New Zealand, and Australia – have published frameworks by which they evaluate the quality of administrative datasets. In some cases, statistical agencies published the results of their evaluations of national data systems. These evaluations range in their level of completeness and in their focus on quantifying the quality of a given dataset.

[29] Schumacher, Britta, *Assessing the Quality of Administrative Data in a Census*, U.S. Census Bureau, (2022), https://www.census.gov/content/dam/Census/programs-surveys/international-programs/stic/STIC%20Assessing%20the%20Quality%20of%20Administrative%20Data%20in%20a%20Census.pdf.

[30] *A Framework for Data Quality,* FCSM, (2020), https://www.fcsm.gov/assets/files/docs/FCSM.20.04_A_Framework_for_Data_Quality.pdf.

[31] *Realizing the Promise of administrative data for enhancing the American Community Survey*, U.S. Census Bureau American Community Survey Office, (2018), https://www.census.gov/content/dam/Census/programs-surveys/acs/methodology/agility-in-action/administrative-records-in-the-american-community-survey.pdf.

[32] *Administrative data research and analysis project (ADRA): Comparisons of data sources on earnings*, U.S. Department of Labor, (n.d.), http://www.dol.gov/agencies/oasp/evaluation/completedstudies/administrative-data-research-and-analysis-project-ADRA-comparisons-of-data-sources-and-earnings.

Priorities for data quality assessment depend on purpose: what research question needs to be answered? Is the goal to generate a trend line or create a statistic? **We identified five priority use cases for analysis of STW training and credentialing using administrative data**:

1. Measuring the rate of attainment of non-degree credentials (NDCs) for STW
2. Measuring aggregate returns to NDCs by credential type
3. Identifying differences by race and gender in the attainment of NDCs
4. Identifying which NDCs are associated with the strongest labor market returns for STW
5. Evaluating the effectiveness of public policies that support the attainment of NDCs

We assessed available datasets on eight dimensions of data quality (see fig. 2). There is no single rubric or instrument for assessing the quality of administrative data, nor is there agreement on whether or how one should quantify the value of a dataset. However, there are standards and guidelines and there is some alignment across these. Across various efforts, we found eight dimensions of data quality that are regularly referenced in quality assessment rubrics for administrative data. We used these eight dimensions[33] to select administrative datasets for review. For 12 of these datasets, we collected metadata, identified relevant variables, and assessed data quality.

Figure 2. Eight Dimensions of Data Quality

[33] Adjustments to the dimensions of data quality assessed included assessment of "coverage" instead of "completeness", addition of "suitability for longitudinal research," and dropping of "accuracy." Accuracy (construct validity, internal validity, and other sources of measurement error as well as consistency of definitions, data collections, and data cleaning/suppression/obscurement over time) was too difficult to assess from metadata and resulting datasets. This dimension should be explored with more time for conversations with data administrators and various actors contributing to data collection and cleaning.

Our data quality assessments covered the above-described aspects of administrative data quality. All datasets were assessed based on publicly available documentation. When possible, we interviewed data producers and conducted our own analysis of the data files (e.g., to identify the percentage of cases with missing data for STW-relevant variables). Our analyses examined the datasets holistically, considering their suitability for research that cuts across all occupations – not just for the STW. A key consideration in our evaluation of each dataset was whether occupation and/or field of study data is sufficiently granular to permit the identification of labor market subpopulations such as the STW.

Timeliness and integrity are two quality metrics that show high compliance across the datasets examined. Most datasets have a high rating in terms of interoperability and non-response rates. RAPIDS demonstrates the highest quality rating and highest degree of fitness for statistics on the STW across the high priority use cases identified, though coverage is limited given the small size of this federal program.

Coverage is not uniform across datasets. All supply-side datasets cover unique, though not necessarily exclusive, sets of participants, training programs, and institutions at different degrees of granularity. Across these datasets, the most common identifiers of the STW are occupational codes following the Standard Occupational Classification (SOC), provided to describe employment (PIRL, RAPIDS, NLx) and to describe training (ETPPR). The PSEO and IPEDS contain Classification of Instructional Program (CIP) codes, which can be cross walked to SOC codes. These codes improve the interoperability of the datasets, allowing us to identify STW-relevant training, STW-relevant employment, or STW individuals, depending on the units of observation in the dataset.

Most of the eight dimensions of data quality were helpful in assessing if the datasets were suited to describing the STW. We were most limited in our ability to assess the accuracy of the datasets and individual variables.[34]

We determined that six national, publicly-available administrative datasets are best suited for research on the STW and STW-relevant training programs and credentials: IPEDS, PSEO, PIRL, RAPIDS, ETPPR, and NLx. These six exhibit high data quality ratings and can be applied to the five priority use cases described above. In the next section, we identify opportunities to describe the STW, with reference to these six datasets.

[34] In the future, accuracy could be assessed in several ways. "Accuracy" may include construct validity, internal validity, reliability, and other sources of measurement error more appropriately addressed when comparative quantitative information is available or a more robust qualitative inquiry could be conducted. A qualitative inquiry could engage various stakeholders to assess the data collection process and the face validity of the results; or it could engage experts to assess the various stages of data production, including data collection and data cleaning processes.

Administrative Datasets and Quality Assessment Results

- WIOA Participant Individual Record Layout (**PIRL**): Full assessment | Executive summary
- Career OneStop Certification Finder: Full assessment | Executive summary
- Career OneStop License Finder: Full assessment | Executive summary
- Army COOL (Credentialing Opportunities Online): Full assessment | Executive summary
- Eligible Training Provider Performance Results (**ETPPR**): Full assessment | Executive summary
- National Labor Exchange Research Hub (**NLx**): Full assessment | Executive summary
- Post-Secondary Employment Outcomes (**PSEO**): Full assessment | Executive summary
- Colorado License Roster (Nursing): Full assessment | Executive summary
- Maryland Noncredit Workforce Completers System: Full assessment | Executive summary
- Credential Engine Credential Registry: Full assessment | Executive summary
- Integrated Post-secondary Education Data System (**IPEDS**): Full assessment | Executive summary
- Registered Apprenticeship Partners Information Data System (**RAPIDS**): Full assessment | Executive summary

Full results can be found in Appendix C, or at https://gwipp.gwu.edu/administrative-data-repository-skilled-technical-workforce.

OPPORTUNITIES TO DESCRIBE THE STW USING ADMINISTRATIVE DATA

The NCSES defines the skilled technical workforce (STW) and skilled technical workers by their STEM skills and lack of a bachelor's degree. A 2019 NSB report operationalizes this definition by identifying occupations requiring STEM skills but with less than 50 percent of employees in that occupation holding a bachelor's degree.[35] This demonstrates it is possible to identify the STW through occupational codes and instructional codes that identify individuals have STEM skills.

Identification of the STW is also possible via information on training, skills, credentials—and employment, based on typical or actual level of training and skills required on the job. Ideally, to identify STW individuals, information on STEM skills is paired with information on highest level of education. With this in mind, we reviewed the results of the data quality assessment to identify six opportunities and potential next steps to describe the STW using administrative data.

The first opportunity is to describe the demographics of participants in training programs relevant to the STW, across publicly available federal administrative datasets. In our quality

[35] The National Science Board defines the Skilled Technical Workforce (STW) as "individuals who utilize science and engineering skills in their jobs but do not have a bachelor's degree." See *The Skilled technical workforce: Crafting America's science & engineering enterprise* (NSB-2019-23), National Science Board, (2019), https://nsf-gov-resources.nsf.gov/nsb/publications/2019/nsb201923.pdf. NSB operationalizes this definition in 2019 by identifying occupations requiring STEM knowledge and skills but with less than 50 percent of employees in that occupation holding a bachelor's degree—"skilled technical workers, defined here as workers in occupations that employ significant levels of S&E expertise and technical knowledge and whose educational attainment is less than a bachelor's degree"—from the "Demographics of the Skilled Technical Workforce" chapter in the September 2019 Science and Engineering Labor Force report at https://ncses.nsf.gov/pubs/nsb20198/the-skilled-technical-workforce#demographics-of-the-skilled-technical-workforce.

assessment, we found few use cases where multiple datasets were an excellent fit, with one exception: multiple administrative datasets contained demographic data to identify demographic trends in STW-relevant NDC attainment. Various datasets with demographic trends contain occupational codes or instructional codes to identify STW-relevant programs.

The available data primarily covers programs at institutions in which the U.S. government is making significant investments—e.g., at educational institutions eligible for Title IV funding and at training institutions eligible for WIOA funding. The coverage is national, and the datasets reviewed can immediately contribute to our understanding of STW-relevant programs. Data is made publicly available by the Department of Labor, the Department of Education, and the U.S. Census Bureau. All the relevant datasets from these agencies have a high level of integrity, timeliness (they lag real time significantly but are regularly updated), granularity, response rate, and potential for interoperability.

The ideal unit of analysis to understand career pathways is the individual worker, and granular information on individual program participants is held by the Department of Labor, responsible for reporting related to WIOA and apprenticeships. But most publicly available information from across agencies is on education and training programs.

A second opportunity is to better identify NDCs across datasets with common identifiers for credentials, credential issuers, training programs, and training organizations. This would improve the relevance of datasets to describing STW training and career paths and their interoperability. This applies to various datasets maintained by the U.S. Department of Labor. For these, inconsistent definitions of the specific credentials attained would have to be addressed; additional data collection and/or crosswalks could help to identify similar NDCs across datasets. In the future, the U.S. Department of Education may also seek to identify NDCs via their voluntary collection of non-credit program information, which could expand the scope of their data collection to professional credentials in addition to academic degrees and certificates.

The most prominent datasets from the U.S. Department of Education National Center for Educational Statistics (NCES) and U.S. Department of Labor Employment and Training Administration (ETA) both track credential attainment but with no available crosswalk for comparison. The information on certificates in NCES IPEDS is detailed and precisely defined, but degrees and certificates are the only credentials tracked, with a new but separate system for voluntary reporting on non-credit programming.[36] DOL ETA PIRL collects information on attainment of a broader spectrum of training and credentials, but only limited details on the programs and credentials themselves could be compared to IPEDS or other data sources (no info on time required, as in IPEDS or ETPPR, or on whether assessment is required, as in

[36] The NCES IPEDS has near comprehensive and uniform coverage of for-credit programming across colleges nationwide and 20+ years of data collected. There are detailed categories for type of certificates awarded and demographic information on program completers. The ETA PIRL, with data available on a quarterly basis dating back to 2017, includes important details on participants in programs funded to serve the unemployed and underemployed, including participant demographics, training program enrollment, and program completion.

Certification Finder from ETA CareerOneStop). The DOL ETA Eligible Training Provider Performance Results (ETPPR) dataset makes important progress compiling information on training programs recognized by state authorities nationwide, and provides some additional details on programs (e.g., duration, pre-requisites), but no additional information on credentials is attained through these programs.

The expansion of non-credit data collection by NCES would generate new information on STW education and training trends and could align definitions with ETA. ETA may also be able to create more continuity in identification of credentials across its various datasets created for programmatic purposes: RAPIDS, ETPPR, and PIRL. Categories currently used in PIRL include: secondary diploma or equivalent, AA diploma/degree, BA diploma/degree, occupational licensure, occupational certificate, occupational certification, other, and none.[37] A new Post Secondary Credential Attainment Tool on the ETA WIOA website[38] is available to determine if a credential meets the WIOA definition of one of the recognized postsecondary credentials.[39] The potential to gain a credential is also reported related to training programs with a similar level of specificity.

Common identifiers for credentials, credential issuers, training programs, and training organizations would better identify NDCs across datasets. This would allow for comparing education and training providers across NCES and ETA datasets for a more complete picture of STW education and training.[40]

A third opportunity is to improve coverage and completeness of federal datasets by addressing state-by-state variation in data quality, an important source of non-random missing data**.** Across existing administrative data collection relevant to the STW, there are types of information that are persistently difficult to collect and/or integrate into statistics, research, or publicly available data products, given missing or problematic self-reported information. Missing or imprecise observations can result from administrative data collection efforts. If these persist, analysts avoid reporting and/or interpreting the results. Too often no reports on missing data are provided at all, not even on the extent of the limitations and risks to interpretation. A culture of transparency could be encouraged, with constructive discussions of how imperfect data and missing observations inhibit effective policymaking at the federal and state level. NCSES could contribute to these discussions by identifying the most important sources of variation in missing or incomplete data.

---

[37] *WIOA Participant Individual Record Layout (PIRL)*, Department of Labor Employment and Training Administration, (n.d.), https://www.dol.gov/sites/dolgov/files/ETA/Performance/pdfs/ETA_9170%20PY%202022%20(Accessible)%20.pdf.

[38] This tool may improve the transparency and consistency of credential reporting. ETA's new dashboard for WIOA programs is helpful for data users interested in participant demographics and will improve as more elements are added. In the meantime, it's important that the definitions and supporting documentation available for program reporting are provided in the technical documentation for researchers.

[39] *Post Secondary Credential Attainment Tool*, U.S. Department of Labor, Employment and Training Administration, (n.d.), https://www.dol.gov/agencies/eta/Performance/resources/credential-attainment.

[40] In two sequential reports, the credential transparency organization Credential Engine has called for improved methods for matching IPEDS' educational institution dataset and ETA's Eligible Training Provider Performance Results (ETPPR) training institution dataset. https://credentialengine.org/wp-content/uploads/2023/01/Final-CountingCredentials_2022.pdf https://credentialengine.org/wp-content/uploads/2021/02/Counting-Credentials-2021.pdf.

We found major limitations, and opportunities for improvement, in the reporting of individual participant demographics as well as in the reporting of detailed training programs and locations. Missing data suggests persistent challenges in data collection and reporting in some states. For example, the PIRL dataset suggests wide variation in the degree of missing observations across states, suggesting room for improvement in the states with the most missing data. In IPEDS, institutions vary in the extent to which they report detail at the campus level, versus aggregate information linked to a headquarter location. Reporting units may also use different CIP codes to report similar programs. PIRL and other ETA datasets vary in the extent to which key geographies, such as workforce board service areas, are identified.

Detailed discussion of missing data and discrepancies are included in Appendix A. The ETA PIRL records and the DOL ETA RAPIDS apprenticeship data collections constitute the most extensive national individual record collection publicly available to researchers and statisticians interested in the STW, including training participant demographics and the location of training.[41] The issues found in these datasets are apparent in the new data collection for ETA ETPPR, which stems from the ETA PIRL and state data collections. Where it appears, employment information also has prevalent missing information, especially if self-reported, but also if sourced from Unemployment Insurance records.

The next section describes four recommendations for NCSES leadership and inter-agency coordination to describe the STW and improve the underlying datasets.

RECOMMENDATIONS FOR NCSES AND RESEARCH PARTNERS

Inter-agency coordination is necessary to improve STW education and training programs. Administrative data can be foundational to these efforts, if federal agencies agree to enhance the interoperability and quality of datasets generated for education and training program monitoring and reporting.

Although administrative datasets have weaknesses such as missing data and incomplete population coverage, with improvements and linkages across datasets, administrative data can be used to fill the gaps inherent to federal statistical surveys. Administrative datasets can provide more granular and detailed information on the demographics, educational backgrounds, and economic activities of members of the STW.

Above we identified three opportunities, which we discuss further below in terms of recommendations: describe the STW using existing federal data sources, expand information on key units of analysis, fill gaps important to public policy priorities. Our fourth

[41] The coverage is much better on the LEHD, but very difficult to identify STW (no occupation codes; PSEO incomplete) and individual records are accessible only via FSRDC.

recommendation below suggests expanding microdata access to collaborating researchers within and outside of government agencies via optimized microdata access and linkages.

Four recommended actions for inter-agency coordination and alignment will enhance collective understanding of the STW and improve STW-relevant programs:

1) **USE THE DATA WE HAVE:** Using existing administrative data in NSB reports can lay the groundwork for longer term efforts necessary to improve the quality of federal administrative data and set standards for future data review.
2) **COLLECT SIMILAR INFORMATION ON KEY UNITS OF ANALYSIS**: Federal agencies funding STW-relevant programs could collect consistent information on program participants, credentials, training programs, and training providers.
3) **FILL DATA GAPS IMPORTANT TO PUBLIC POLICY**: Agencies with large datasets could invite researchers to understand and reduce missing data on the location and demographics of individuals and training programs.
4) **OPTIMIZE MICRODATA ACCESS and LINKAGES:** NCSES and federal statistical agencies can invite researchers to access and link microdata for the purpose of describing STW career pathways through mechanisms such as the America's Data Hub Consortium and the NCSES Data Enclave.

Below we discuss the next potential steps for each recommendation.

Use The Data We Have

We focused our data quality review on information regarding NDCs relevant to understanding the STW. We found that demographic information was available across many federal agency datasets for groups participating in STW-relevant programs receiving significant national investment. These datasets are not representative of the entire STW with the truly national coverage of the NTEWS, but each provides unique, granular insights into training and credential trends for a specific population or credential type.

The NCSES can include select information from available administrative datasets in national reports, highlighting the unique, granular insights available for specific groups participating in public programs. This could demonstrate the value of administrative data for understanding the STW and help us understand key STW populations. Since demographic information is available from both federal administrative datasets and statistical survey results, describing STW demographics may be a useful place to start in evaluating insights that can be gained from administrative data.

While individual-level de-identified data sets that are longitudinal are the most valuable for understanding the career pathways of the STW, these are difficult to aggregate across states. There is currently more information on education and training programs readily available and linked to demographic information. This information on STW-relevant education and training programs could be summarized in NCSES and NSB reports, describing all STW-relevant higher

education for-credit programs and all WIOA-funded training programs. This information on training programs from Education and Labor agencies include topics, locations, aggregate participant information, and associated credentials (though detail on credentials in the WIOA dataset is limited).

Collect Similar Information on Key Units of Analysis

Federal agencies can aim to track consistent information on training and credentialing across data collection efforts relevant to the STW. More alignment across Education and Labor data collection efforts would improve the interoperability of the datasets and could help to set standards for emerging and still nascent data collection at Energy, Commerce, Health and Human Services programs. Labor and Education have different approaches to data collection and publication that could be aligned with respect to key units of analysis.

There are several important units of analysis that can be described using existing data, including individuals (students/workers), credentials, training programs, and training providers. To improve the interoperability of datasets created for programmatic and performance evaluation purposes, federal agencies could agree on a common core set of information. For training program, this includes duration of program, program delivery format, training provider, and the organization awarding any associated credential. For training providers, opportunities exist to better track their location, the organization offering the training, and the associated credentials. For individuals participating in programs, opportunities exist to better track their location, age, race and ethnicity, highest level of education, progress through programs including enrollment and completion, and employment results.

Federal agencies will need to adjust their existing data collection and management systems to align with a "common core" record layout. Other data owners, such as those that issue certifications in the non-profit and private sector or state licensing authorities, may be motivated to do so as well if incentives and technical support are available.

Fill Data Gaps Important to Public Policy

Missing data is a common feature of administrative data—studies have shown its prevalence in the management of healthcare administrative data, revealing limited and varied levels of accuracy that raise concern over its validity. Nonetheless, this data can be used for decision making. A November 2022 Executive Order directed at Office of Management and Budget (OMB) noted that gathering data is the first step.[42] For example, a collaboration between the Department of Labor and the Department of Commerce could lead to demographically disaggregated information on employment results from subgrantees of Commerce programs.

---

[42] Executive Order. See, *Announcing November 29. 2022 open government engagement session on increasing federal data access and utility | ostp.* The White House, (2022), https://bidenwhitehouse.archives.gov/ostp/news-updates/2022/11/22/announcing-november-29-2022-open-government-engagement-session-on-increasing-federal-data-access-and-utility/#:~:text=November%2022%2C%202022-,Announcing%20November%2029%2C%202022%20Open%20Government%20Engagement%20Session%20on,Federal%20Data%20Access%20and%20Utility&text=The%20Biden%2DHarris%20Administration%20is,government%20data%20for%20all%20Americans.

Addressing missing data for the specific purpose of describing the STW can improve buy-in across federal agencies. All agencies funding STW-relevant programs[43] could better articulate what we understand about the technical and human processes that result in missing administrative data. With enhanced ability to explain gaps in knowledge, such as in the context of the STW training profile, administrators and policymakers can prioritize improvements. For example, a broader understanding of discrepancies in wage reporting could increase awareness of limitations and enhance confidence in the data available, while targeting resources toward addressing key missing records. Discrepancies regarding which programs are offered in which locations could be addressed in tandem. Enhancing both wage reporting and location specificity could improve our understanding of STW career pathways and employment outcomes in the context of real labor markets and regional economies.

Optimize Microdata Access and Linkages

Researchers, scholars, and statisticians can help to examine coverage and variation by states from which administrative data is sourced. These collaborators could also help to augment capacity in states with fewer data collection resources and data quality control processes. Differences, if not addressed, will hinder analysis of existing data and limit new data collection on non-degree and non-credit programs.[44] Expanded access to microdata could help researchers to provide new methods and insights, especially if individual records and other important units of analysis could be linked across datasets.

This increased state capacity could vastly improve results at the federal, state, and local level. State Longitudinal Data Systems, which link K-12, college, and employment data, could enhance federal data collection on the STW if partnerships with states are strengthened. Stronger state research programs, such as those facilitated by the Multi-State Data Collaborative, the Coleridge Institute, and National Association of State Workforce Agencies (NASWA), could improve the SLDS.[45]

To unlock the power of administrative data, linkages across datasets are key. Linking datasets requires improved methods and may require training. Data security expectations for partners seeking access to microdata could be more reasonable and transparent to encourage participation so that relevant capabilities can be developed and enhanced over time.

---

[43] In May 2023, at least 38 federal agency offices or departments administered programs relevant to the STW. See https://bpb-us-e1.wpmucdn.com/blogs.gwu.edu/dist/b/3597/files/2023/06/Skilled-Technical-Workforce-Programs.pdf

[44] Various research efforts have aimed to understand non-credit education and training. See, Xu, D., & Ran, X, *Noncredit education in community college: Students, course enrollments, and academic outcomes*, CCRC Teachers College, Columbia University, (2015), https://www.luminafoundation.org/files/resources/noncredit-ed-in-community-college.pdf and related efforts. See, *State noncredit data publications*, Rutgers University, (n.d.), https://sites.rutgers.edu/state-noncredit-data/publications/ https://www.cael.org/news-and-resources/cael-releases-research-report-on-short-term-high-value-credentials, *CAEL releases research report on short-term, high-value credentials,* CAEL, (2023), https://upcea.edu/creating-non-credit-to-credit-pathways/.

[45] See Schneider, Mark. "What's on Tap at IES for the next Year?." Institute of Education Sciences, (2023), https://ies.ed.gov/director/remarks/04-19-2023.asp on SLDS v.2

Direct access to data for researchers such as through the Standard Application Process (SAP)[46] would accelerate progress. Access to microdata from administrative datasets could be facilitated through this platform. The platform could also indicate if any individual products listed could be linked using common identifiers.

MICRODATA ACCESS OPTIONS

In addition to asking statisticians and researchers to apply through the SAP, several options are available to public agencies and other data owners who want to collaborate and expand secure data access. One model is the creation of **agency or organization-specific data enclaves** that allow researchers to access datasets in a "locked down" remote desktop environment where the data is analyzed on remote systems not connected to the Internet. Such an enclave is used by NCSES to permit access to restricted-use survey files to researchers not inclined to obtain access through a Federal Statistical Research Data Center (FSRDC). In the NCSES data enclave, researchers prepare their analysis independently before moving output files to a review platform through which an NCSES contractor confirms that PII is not inadvertently disclosed in the publication of results.

A second model is generating datasets that have lower risk of disclosure. This includes the **addition of random "noise" to key variables** that could be used to identify individuals. Under this approach, key quantitative variables are randomized slightly to prevent individual records from being identified directly from the microdata. For example, a salary might be changed by a random number, resulting in a number reported in the data file that is up to $100 greater or less than one's actual salary, to prevent matching on the exact amount of one's salary. This approach was recently introduced to the PIRL data files posted online by ETA, and thus far does not appear to have had any meaningful impact on the analyses that can be conducted by external researchers.

A third model is the creation of **publicly accessible synthetic datasets** that closely mirror the attributes of "real" restricted-use files. This approach has been piloted by the Census Bureau to enable the analysis of variables from the Survey of Income and Program Participation (SIPP) that would normally be placed in restricted use files. Synthetic microdata files are tested by agency statisticians prior to public release to ensure that significant relationships that show up in the synthetic files are largely the same as those in the actual microdata – and, in the case of the SIPP, agency statisticians have offered to run code produced for the synthetic files on the actual restricted use microdata to confirm the validity of results from the synthetic files.

Finally, restricted use microdata access can be held by **multi-organization or multi-agency secure platforms**. The Coleridge Initiative and its secure Administrative Data Research Facility is an example of a third party that houses restricted use files produced by state agencies, including state longitudinal data systems. While Coleridge does not currently host federally-produced restricted use files, another emerging initiative, the National Secure Data Service

[46] A recent innovation in data access was the creation in 2022 of the Standard Application Process for researcher access to protected/confidential data across 16 federal agencies through researchdatagov.gov. See *Memorandum for Heads of Executives Departments and Agencies* (M-23–04), Office of Management and Budget, (2022), https://www.whitehouse.gov/wp-content/uploads/2022/12/M-23-04.pdf.

(NSDS)[47], promises to facilitate access to restricted use data from multiple federal agencies. The SAP provides a common application procedure that external researchers can use to propose studies drawing upon microdata from multiple agencies.

WHERE TO START: LINKING FEDERAL ADMINISTRATIVE DATASETS

To improve accessibility and interoperability, it is important to enable microdata access and linkages across federal datasets. Few statistical and/or research organizations have published information on de-identification and matching processes that ensure that data can be linked while maintaining the confidentiality of the records, but this information is increasingly available.[48] These processes are critical to connect information across education, training, and employment systems and to maintain the confidentiality of program participants' personal information. Federal agencies in the U.S. can learn from each other as each builds capabilities to enhance data sharing.

There is immediate potential to link administrative datasets, especially since several are administered by ETA. ETA's ETPPR relies on the same de-identified individual level detailed data that is available for public purpose via the PIRL. ETA ETPPR, ETA PIRL, ETA RAPIDS all result from compilations of state-level data collections, suggesting linkages require state agency cooperation.

IPEDS is an NCES dataset that relies on individual data owned by reporting units (universities and other education providers). Reporting units provide information aggregated at the program level to NCES. Generating individual participant unit records at the national level would require an entirely different reporting process. In this context, two separate processes have emerged. One is that the National Student Clearinghouse collects individual records from a similar population of education institutions and connects these across institutions and across states. The Census Bureau also coordinates linking the data from the same reporting units through an entirely separate process, combining individual data with employment records resulting in a new public use dataset, the PSEO. It will be important for NCSES and partners to understand what we have learned from these various efforts and if there are opportunities for linkages to improve interoperability and access.

WHERE TO INNOVATE: UTILIZING PRIVATE DATASETS AND STATE SLDS

Looking ahead, NCSES could explore sources of data that originate in the private sector as supplemental sources of data to support research on the STW. One of the most promising sources of data – if access can be secured by qualified researchers – is the treasure-trove of self-reported career and credential data held by social media platforms. LinkedIn is of

---

[47] National Secure Data Service (NSDS) was first proposed by the bipartisan Commission on Evidence-Based Policymaking's Final Report in 2017. The Advisory Committee on Data for Evidence Building, established by the Foundations for Evidence-Based Policymaking Act of 2018, provided specific recommendations to refine NSDS in 2022. The National Secure Data Service Demonstration (NSDS-D) project, required under the 2022 CHIPS and Science Act, aims to strengthen data linkage and data access infrastructure at all levels of government and in all sectors to enhance evidence-based decision-making.

[48] NCSES collected information on strategies for maintaining confidentiality of individual records and a National Secure Data Service Demonstration project launched in 2022. See, Plimpton, S., Agency information collection activities: Comment request. *Federal Register*, 65611-65613, (2022), https://www.federalregister.gov/documents/2022/10/31/2022-23629/agency-information-collection-activities-comment-request; *The National secure data service demonstration project*, NSF National Center for Science and Engineering Statistics, (n.d.), https://ncses.nsf.gov/about/national-secure-data-service-demo.

greatest relevance to research on credentials and the STW. Information is robust for some fields and industries and limited in others. Other sources exist that may be of interest to researchers, such as data that is self-reported to job information transparency websites such as Payscale. Records held by the National Student Clearinghouse may also be of value for studying educational attainment and enrollment in programs relevant to the STW. The National Student Clearinghouse's dataset is particularly noteworthy for its coverage of periods of college enrollment that did not result in a degree. Some institutions are also reporting enhanced records to the Clearinghouse in recent years on course enrollment and fields of study, which could provide valuable insights on non-traditional educational pathways into STW occupations.

State Longitudinal Data Systems (SLDSs) are another area of promise for further research on the STW. SLDSs were largely excluded from our project due to their sub-national nature, but a picture of STW educational attainment and career outcomes could potentially be assembled by linking together data from SLDSs in multiple states. A major limitation is the time-consuming process of justifying research to each state agency involved in managing SLDSs; however, this disadvantage is offset with most states providing access without charging a fee.

Looking ahead to new emerging datasets: with more competency- and skills-based programs and platforms, we anticipate the creation of new administrative datasets. These may be linked to data already being collected or held within various government agencies and private-sector organizations. For example, the U.S. Department of Defense maintains detailed transcripts of training for all military personnel as well as information on credentials attained. Similarly, emerging data from competency-based education and training programs could augment our understanding of what a Career and Technical Education (CTE) diploma or apprenticeship completion certificate represents.

CONCLUSION AND DISCUSSION

Recognizing the important contributions of the Skilled Technical Workforce (STW) to the nation's security, innovation systems, and global competitiveness, the National Science Board and National Science Foundation are seeking to expand and upskill this critical workforce segment. However, little is known about what training and credentials are most prevalent and effective for helping individuals enter and excel in STW occupations.

Administrative data has the potential to fill important information gaps about the STW. For example, training participation records collected for programmatic purposes can complement statistical surveys, allowing for detailed analysis of training trends for small geographic areas.[49] Statisticians have experience in using and improving existing administrative and survey data to inform public policymaking. Demonstrating these

[49] See GWIPP-CREC working paper for this project on Administrative Data and the NTEWS

capabilities to describe the STW would be an important public service and public education opportunity that would enhance data-driven decision making.

The National Center for Science and Engineering Statistics (NCSES) is uniquely positioned to bridge siloed data systems given the agency's role as lead administrator of the new National Training, Education, and Workforce Survey (NTEWS) and its broader STW research agenda. In advancing inter-agency coordination to understand the STW, NCSES has a dual role: on the one hand, generating more accurate and granular insights and statistics, and, in so doing, improving administrative data collections across education, training, and employment systems. Increasing the degree of comparability and interoperability across federal data collection efforts will be critical to improving the quality of the information and building confidence in the results for policymaking.

As with survey data, transparency and understanding of limitations and appropriate applications is key to coherent research and insights from administrative data. NCSES can utilize and improve upon our approach to data quality assessment in several ways. Our initial review and discovery process focused on the quality of data in its published format, a *product*-oriented review. However, a *process*-oriented review will be critical to identifying and addressing data quality issues. A process-oriented review involves the analysis of the procedures followed when creating a dataset, rather than a focus on the finished product. The process approach involves collecting data directly from the producers of data, for example by observing as data is collected in the field or interviewing data producers to learn more about data collection procedures and verify that best practices were followed. Most data quality assessments consider both process-oriented and product-oriented elements of quality, and a wide range of language is used to describe different dimensions of quality. "Process" is defined as a set of dimensions (in the authors' words, a "hyperdimension") in a working paper published in 2009 by Statistics Netherlands (Dass et. al. 2009).[50]

Additionally, our search for relevant datasets focused on those that currently include information on NDCs, not all datasets with broader population coverage that could potentially augment the new NTEWS. With the results from the NTEWS available in 2025, NCSES and collaborators can tackle more precise fitness-for-purpose issues related to augmenting specific survey questions, survey response, and resulting statistics.

[50] Daas, Piet, Saskia Ossen, Rachel Vis-Visschers, and Judit Arends-Tóth. 2009. *Checklist for the Quality Evaluation of Administrative Data Sources*. Statistics Netherlands Discussion Paper 09042. The Hague/Heerlen: Statistics Netherlands. Retrieved October 15, 2023 (https://ec.europa.eu/eurostat/documents/64157/4374310/45-Checklist-quality-evaluation-administrative-data-sources-2009.pdf/)

APPENDIX A: DETAILED OPPORTUNITIES FOR IMPROVEMENT OF INDIVIDUAL DATASETS

Typical issues faced by researchers conducting analyses of administrative data include developing conceptual frameworks, defining the population covered, identifying the unit of analysis, ensuring that the "sample" size is large enough, identifying coverage gaps due to unit and item nonresponse, and assessing measurement error. Other issues with using this data for statistical purposes include issues with programmatic continuity, data collection process changes, changes in variable meaning over time, privacy concerns where there are detailed records on individual participation, and potentially costly data improvement and transformation activities to prepare the data for research purposes.[51]

We found many of these issues in the administrative datasets reviewed for this paper. However, these issues are not permanent obstacles to integrating administrative data into research and statistics. Administrative data can provide granular information on key subsets of the STW population for real labor markets and other key economic geographies. Data quality issues and the viability of producing statistics should be evaluated in the context of specific research questions, regarding the reliability, timeliness, granularity, accuracy, and other dimensions of data quality necessary for producing the relevant statistics. With results from the NTEWS in 2024, the value-added contribution of administrative data can also be assessed in the context of understanding survey responses and improving response rates for the new survey.

The six datasets selected to inform recommendations to NCSES were chosen due to their integrity, relevance, accessibility, potential interoperability, and other dimensions of quality, as well as their fit for priority use cases. They include information on training program participants and employment outcomes (PSEO, ETPPR, PIRL), training program participants and employers (RAPIDS), education program completion and demographics (IPEDS), and education/training expectations by employers (NLx).

The dataset-specific recommendations below address various data quality issues, with a focus on relevance, coverage and timeliness, and access. These observations informed our identification of opportunities to describe the STW using administrative data and the four recommendations for inter-agency alignment in this report.

POSTSECONDARY EMPLOYMENT OUTCOMES (PSEO) – U.S. CENSUS BUREAU

The PSEO is a breakthrough project that connects data across educational institutions, and across states, to produce information on the employment outcomes of students completing

[51] See Burwell, S., *Memorandum for the heads of executive departments and agencies*, Office of Management and Budget, (2014). https://obamawhitehouse.archives.gov/sites/default/files/omb/memoranda/2014/m-14-06.pdf and the MIT Handbook on Using Administrative Data for Research and Evidence-Based Policy https://admindatahandbook.mit.edu/.

degrees and certificates. The data for this product is sourced from institutions that are accustomed to providing information on their degree and certificate completers to the National Center for Education Statistics. We identify the following opportunities for improving PSEO:

- **Relevance**
  - Integrate data on the race, gender, and other demographics of graduates available to help identify differences by race and gender in the attainment of NDCs
  - Investigate options to capture information about employment for those currently excluded, who are unemployed or marginally employed after graduation.
  - Investigate options to document previous education, previous or concurrent employment, and concurrent degree and certificate attainment.
- **Coverage and Timeliness**
  - Continue to recruit additional states and institutions to expand coverage.
  - Clarify the update schedule in the technical documentation or set one if not already in motion.
- **Access**
  - Establish a researcher access point that expands access to the microdata for the purpose of understanding certificate attainment and earnings for the STW in economic regions nationwide for place-based and demographic analysis.
  - Facilitate easier data linkage on individual cases by offering secure access to unique identifiers in microdata available to researchers as well as to state agencies and the educational institutions contributing their data to make this product possible.

ELIGIBLE TRAINING PROVIDER PERFORMANCE RESULTS (ETPPR) – U.S. EMPLOYMENT AND TRAINING ADMINISTRATION

The compilation of state ETPPR into a national searchable dataset is a breakthrough that increases public access to information on available training programs. The dataset and data dictionary are available for download at trainingproviderresults.gov. We identify the following opportunities for improving ETPPR:

- **Relevance**
  - Enable reporting of multiple credentials associated with training programs.
  - Integrate data on the race, gender, and demographics of those completing these training programs to help identify differences by race and gender attainment of NDCs where sufficient sample size exists.
- **Coverage and Completeness**
  - Set and promote a plan to reduce the rate of missing data, focusing on where there are differences across states.
  - Identify and communicate the causes of the missing data. For example, if issues are known, add a field to describe if missing data is due to suppression or non-reporting or other reporting conditions in progress (e.g., program just launched).
- **Access**

- Publish technical documentation on federal data collection and cleaning processes.
- Encourage states to publish information on their own data collection, cleaning, and submission processes.
- Establish a standard reporting/update cycle.
- Provide guidance on identifying outliers and potential errors.
- Facilitate matching training providers across state and federal data collection processes by providing a unique ID to training providers that would enable linking ETA ETPPR and ETA PIRL with NCES IPEDS institutions.

REGISTERED APPRENTICESHIP PARTNERS INFORMATION DATABASE SYSTEM (RAPIDS) – U.S. EMPLOYMENT AND TRAINING ADMINISTRATION

This historical database includes information on apprentices, employers, and their training programs. Only one STW credential is represented (all programs are associated with a Certificate of Completion from the U.S. Department of Labor awarded in collaboration with states and employers). Two critical classification systems are integrated documenting both occupations for training and the industries of the sponsoring employers (SOC and NAICS). A new dashboard makes more information publicly available, including the demographics of participants. We identify the following opportunities for improving RAPIDS:

- **Relevance**
  - Collect and publish information on Related Training Instruction providers.
  - Collect information on relevant NDCs required to start or complete apprenticeships.
  - Include the Total Apprentices Served table to show a more comprehensive picture of the Registered Apprenticeship (RA) system activity.
- **Coverage and Completeness**
  - Continue to recruit additional states to expand coverage.
  - Set a research agenda to reduce the rate of missing data for demographics and add information on Related Technical Instruction provider, instruction duration and intensity.
  - Explain why there are zero or negative values and, if possible, reduce zero or negative values in variables for term length, wage, active apprentice count, and workforce size.
  - Consider if there are opportunities to improve the reliability of data by clarifying and aligning expectations across Office of Apprenticeship (OA) and State Apprenticeship Agency (SAA) states.
- **Access**
  - Make the data request process explicit.
  - Create technical documentation of dataset construction – data collection, cleaning, publishing processes with variation by state

- Define the categories for the education level, program type, organization type and RTI length type variables in the data dictionary and update the data dictionary as the variables in the dataset change or adapt
- Continue to increase transparency and availability of information.
- Collaborate with state agencies to study participants' earnings after exiting an apprenticeship program; take advantage of existing federal and state data matching programs
- Maintain confidentiality while providing researchers access to personally identifiable information that could help link the data to other datasets, include to other datasets at Department of Labor ETA

PARTICIPANT INDIVIDUAL RECORD LAYOUT (PIRL) – U.S. EMPLOYMENT AND TRAINING ADMINISTRATION

This dataset includes de-identified individual records of participation in public programming, including participant demographics and wages earned before and after participation. We identify the following opportunities for improving the PIRL:

- **Relevance**
  - Provide definitions for credential attainment categories.
  - Provide guidance to researchers to identify unique individuals within and across years.
  - Include the proximate locations of individuals, e.g. zip codes or counties.
  - Improve the granularity of earnings data by moving from quarterly to monthly reporting, including flags for workers who gain or lose employment in the middle of a reporting period.
  - Collect additional details related to the type and date of an individual's credential attainment and owner/offeror of credential
- **Coverage and Completeness**
  - Continue to provide training to data providers/states to increase data quality and response rates.
  - Ensure equitable and fair reporting standards by having standardized definitions of success rather than state-specific definitions to allow for more accurate comparison between states.
- **Access:**
  - Increase potential for state agencies and research partners to vet the accuracy of data reported
  - Make the data more accessible by offering smaller downloadable files or API query
  - Maintain confidentiality while providing researchers access to personally identifiable information that could help link the data to other datasets, include to other datasets at Department of Labor ETA

NATIONAL LABOR EXCHANGE (NLX) - NATIONAL ASSOCIATION OF STATE WORKFORCE AGENCIES

The NLx is a new database of job postings that represent employer expectations for hiring including relevant training and credentialing. NLx administrators are exemplifying good administrative data practices with transparent processes for data cleaning and improvement, with a data dictionary published that describes variables for occupation and education. We identify the following opportunities for improving NLx:

- **Relevance**
  - Prioritize and align action to identify credentials and training for the STW by funding collaborative projects that produce public goods on defined topics of education and credentialing, in order to advance structured data available for statistical purposes
  - More explicitly partner with and engage federal and state agencies with capacity to advance data quality while building capacity across state agencies.
  - Categorization of companies by NAICS code could be added along with a parent-child relationship introduced to link firms to their establishments. NLx Research Hub can leverage state LMI agencies' knowledge and experience to link data to verify company identity.

- **Coverage and Completeness**
  - Benchmark progress and create a rating to assess the maturity of variables to inform insights that can be extracted from the database
  - Share current methods for classifying education and occupations so these can be enhanced by researchers and scholars participating in the Research Hub.
  - Fill in missing data on education and occupations through partnerships by increasing data availability, as described below.

- **Access**
  - Promote Data Hub access to a broader array of organizations both public and private
  - Provide clear direction for users to access the data once their application is accepted.
  - Maximize the public benefits of broader participation by linking data access privileges to community contributions, following an open-source model, including requirements to post relevant coding programs for API call and data analysis

INTEGRATED POSTSECONDARY EDUCATION DATA SYSTEM (IPEDS) – U.S. DEPARTMENT OF LABOR

IPEDS is an annual collection of 12 related survey components that gathers data from every college, university, and technical and vocational institution that participates in federal student financial aid programs. The data can be found at www.nces.ed.gov/ipeds. We identify the following opportunities for improving IPEDS:

- **Relevance**
    - More consistent tracking of the parent/child relationship between institutions and their programs at various locations, differentiating between flagship or headquarter and branch or regional locations.
    - More detailed information on family income would be helpful to researchers. The only information available is on the percentage of students receiving financial aid at the institution.
    - More information on educational history for program completers and any overlapping or concurrent credential attainment information.
    - More information on which programs are offered online or only online.
    - Emerging issues include overlap with new credentialing systems (e.g., badges) and delivery platforms (e.g., MOOCs).

- **Coverage and Completeness**
    - Aim to identify non-credit, non-Title IV institutions, and all non-Title IV programs at Title IV institutions.
    - There are no missing values that we identified.
- **Access**
    - Existing platforms have been stable and should be maintained.
    - Differences in the way that data is submitted and how it is presented via public data products could be more transparent, specifically regarding how duration is reported and categorized (e.g., certificates are reported in weeks but published in terms of credit hours.)

APPENDIX B: METHODOLOGY

In collaboration, George Washington Institute for Public Policymaking (GWIPP) and the Center for Regional Economic Competitiveness (CREC) conducted an analysis of administrative data quality through several tasks between 2021 and 2023.

Metadata Results

CREC prepared an inventory of non-degree credential (NDC) datasets. The steps involved adoption of criteria by which to select NDC datasets (including the quality, usefulness – whether it includes relevance to major research questions, interoperability with other data sources, access, and whether it is open for feedback, among others); Identification of 14 NDC auxiliary datasets for repository (PIRL, PSEO, RAPIDS, Certification Finder, License Finder, Burning Glass Technologies, License Rosters, Eligible Training Provider List, NLx, Maryland Noncredit Workforce Completers System, IPEDS, COOL, Course Report, WEAMS public); documentation of datasets that will enable quality assessments; and design of initial NDC variables list including 800 potentially relevant variables across datasets.[52]

Table B1. Variable categories and number of variables within each category across datasets

Variable Category from Admin Data	Total Variables in Admin Data Category
Credential, Skill, Service, or Experience Gained	229
Demographics and Other Participant Background Information	213
Output from Credential, Skill, Service, or Experience Gained	152
Employment and Wages	127
Classification of Industry, Occupation, Institution, Employer/Sponsor/Program	83
Uncategorized	58

Data Quality Assessment

GWIPP assessed the quality of select datasets for research purposes considering both process-oriented and outcomes-oriented approaches. GWIPP established the steps involved to conduct the quality assessments for 12 administrative datasets (Certification Finder, IPEDS, License Finder, License Rosters, NLx, National Student Clearinghouse (NSC), PIRL, PSEO, RAPIDS, ETPPR, and Maryland Higher Education Completers System); specified the framework for assessing the quality of the datasets (relevance, coverage, granularity, timeliness, integrity, accessibility, interoperability, sustainability for longitudinal research, and consistency); and finally, completed the data quality assessments. Of these datasets, some had a focus on attainment and outcomes (IPEDS, PIRL, PSEO, ETPPR, and RAPIDS), while some offered more information on credentials (NLx, Certification Finder, and License Finder).

---

[52] *STW non-degree credentials metadata repository*, GW Institute of Public Policy, (n.d.), https://gwipp.gwu.edu/stw-non-degree-credentials-metadata-repository.

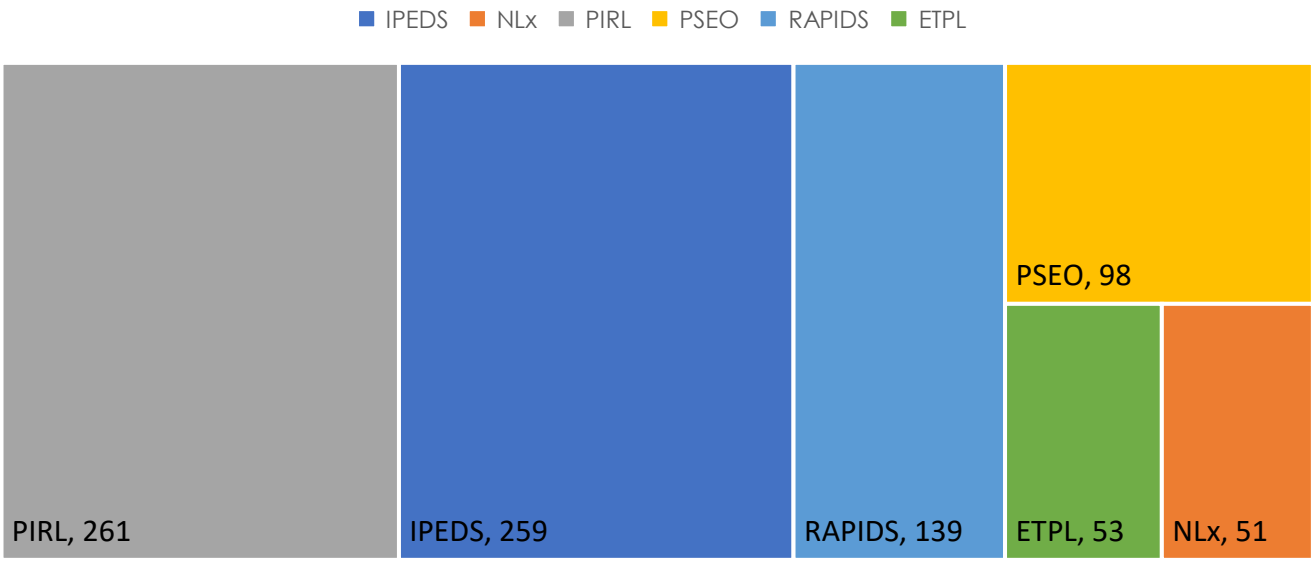
Each dataset subject to the data quality assessment was evaluated using a thorough rubric addressing the relevance of the dataset to NDCs, the population coverage, how granular the data is (how many different categories exist) for key variables of interest (attainment, field of study, income), timeliness of the dataset in terms of how often it is updated and how often data is collected, the integrity of the data and any potential risks to integrity, ease and cost of access to researchers or potential variables withheld, interoperability or the presence of unique identifiers, sustainability for longitudinal research and consistency of key variables over time, among other criteria.

Some of the biggest issues encountered consistently throughout the datasets were lack of well documented technical documentation, lack of data on demographics (age, gender, race/ethnicity data), coverage by state and provider, other missing data, and information lacking such as cost of credentials and earnings before and after obtaining a credential.

Developing Recommendations

Following the GW quality assessments on 12 datasets, we selected six datasets (PIRL, PSEO, ETPPR, NLx, and RAPIDS, and IPEDS) to inform opportunities for describing the STW and recommendations for inter-agency alignment.

Figure B1. Number of variables for selected datasets



CREC interviewed data users and data producers to inform recommendations. CREC summarized the results from each quality assessment and conducted interviews with producers of the administrative data at different federal agencies including the Department of Labor's Employment and Training Administration (ETA) and the U.S. Census Bureau. The research team interviewed scholars who have utilized these administrative datasets in their research – Daniel Kuehn of Urban Institute for RAPIDS data, Randall W. Eberts at the Upjohn

Institute for ETPPR and PIRL, Vicki Lancaster at the University of Virginia's Biocomplexity Institute and Initiative for NLx, William Mabe at Rutgers for IPEDS, and Peter Riley Bahr at University of Michigan for IPEDS. These interviews helped to prioritize our recommendations for each dataset as well as identify areas for further research.

For each dataset, the use cases we considered while developing recommendations were the following:

1. Measuring the rate of attainment of non-degree credentials (NDCs) for STW
2. Measuring aggregate returns to NDCs by credential type
3. Identifying differences by race and gender in the attainment of NDCs
4. Identifying which NDCs are associated with the strongest labor market returns for STW
5. Evaluating the effectiveness of public policies that support the attainment of NDCs

Table B3 provides an overview of the relevant content from each dataset.

Table B2. STW-related administrative data informing recommendations

Agency or Organization	Admin Data Program	Certifications	Licenses	Certificates	Training or Work Experience	Previous Educational Attainment	Occupation	Representation of STW
National Association of State Workforce Agencies	**National Labor Exchange (NLx)**	TBD	TBD	TBD	TBD	TBD	SOC codes	Relevant jobs and job details
U.S. Department of Education	**IPEDS**	No	No	Yes	Yes	No	CIP codes	Relevant degree and certificate programs and participant demographics
U.S. Department of Labor	**Participant Individual Record Layout (PIRL)**	Yes	Yes	Yes	Yes	Yes, including "non-degree certificate"	SOC and CIP codes	Individuals seeking unemployment /services, and relevant training
	Eligible Training Provider Performance Results, TrainingProvider Results.gov	No	No	No	Yes	No	SOC and CIP codes	Relevant training programs and participant demographics
	RAPIDS	No	No	Yes – for journeyman	Yes – for apprenticeships	No	SOC and NAICS codes	Relevant training programs and participant demographics
U.S. Census	**LEHD PSEO**	No	No	Yes	Yes	No	CIP and NAICS codes	Relevant degree and certificate programs

Source: CREC, based on template used by National Science Foundation for summary of statistical survey content

APPENDIX C: ADMINISTRATIVE DATASET QUALITY ASSESSMENT RESULTS

This appendix reproduces the Administrative Data Quality Assessments (and their Executive Summaries) produced by the George Washington University's Institute for Public Policy (GWIPP) and the Center for Regional Economic Competitiveness (CREC). These can also be found online at https://gwipp.gwu.edu/administrative-data-repository-skilled-technical-workforce.

Use the links in the table of contents below to navigate to the summary and assessment for each dataset:

Contents

WIOA PIRL (Workforce Innovation and Opportunity Act, Participant Individual Record Layout)

EXECUTIVE SUMMARY

Dataset Name: **WIOA PIRL (Workforce Innovation and Opportunity Act, Participant Individual Record Layout)**

Publisher: Employment and Training Administration, U.S. Department of Labor

Website: https://www.dol.gov/agencies/eta/performance/reporting

Unit of Analysis: Individual

Purpose of Dataset: The WIOA PIRL is used to calculate performance statistics about the U.S. public workforce system. It aggregates data collected by individual states on individuals who participate in any type of WIOA workforce development program, including programs for displaced workers, youth, and disabled adults. It includes data on post-program labor market outcomes taken from state unemployment insurance wage records, which is intended to be used to evaluate the effectiveness of WIOA programs.

Quality Metrics

Metric	Rating	Summary Explanation
Non-response	Excellent	With 20,356,612 records in the most recently released data file (Q1 2021), we believe that the dataset contains records on all Americans who have participated in or completed a WIOA program within the year prior to the reference period.
Coverage	Fair	Due to computing power limitations, we were unable to calculate exact coverage rates but based on prior research with the PIRL we note that there are significant issues with missing data. May fields are blank by design: for example, there are only about 700,000 cases with data to report on credential attainment because the remainder of the population did not attain a credential with WIOA support. Missing data for certain fields seems to be concentrated in specific states.
Granularity	Excellent	The PIRL contains detailed (though sometimes top-coded) data on pre-entry and post-completion earnings. It also contains detailed information on the credentials completed while receiving WIOA support, including text strings indicating the full name of the training provider.
Consistency	N/A	We have not been able to find datasets that would lend themselves to an "apples to apples" comparison with NSC.
Timeliness	Good	Data tends to be released to the public with a lag of about one year, though this lag has been decreasing recently. At one point during the COVID-19 pandemic the lag time was two years.
Integrity	Excellent	We did not identify any risks to data integrity.
Accessibility	Fair	Anyone can download the data from the ETA website, however the file is massive and requires significant computer capacity to do even simple calculations.
Interoperability	Good	One should be able to link to IPEDS or TrainingProviderResults.gov at the training program level on the basis of the name and location of the training provider. However, such interoperability could be improved with the creation of a unique common identifier for training providers. Since most identifying personal information is suppressed to maintain confidentiality, it would likely not be possible to make linkages at the individual level to other datasets.
Suitability for Longitudinal Research	Good	Data files go back to 2016 and can be freely downloaded from the DOL website, though computing power may be an issue for analyses that integrate multiple files.
Overall Recommendation	Good	While limited to the WIOA population, this dataset contains rich data on the pre-and post-completion experiences of individuals who obtain many different types of non-degree credentials.

Relevance to Use Cases

Use Case	Rating	Summary Explanation
Analyze the Overall Prevalence of NDCs	Poor	The WIOA PIRL only covers less than 10 percent of the U.S. adult population.
Identify Which NDCs are Associated with Highest Earnings	Good	Detailed analyses relating the attainment of different credentials to earnings can be completed with the PIRL, keeping in mind the limitations of the nature of the population.
Identify Patterns of Inequality in NDC Attainment	Excellent	In addition to detailed labor market data, PIRL contains indicators of demographic characteristics.

Enrichment of NTEWS Microdata	Good	It is possible that restricted versions of the PIRL may contain identifiers that would permit linkages with restricted versions of the NTEWS at the individual level for at least some subset of the skilled technical workforce. Moreover, one could link at the level of occupation or field of study to compare labor market outcomes reported in the PIRL and in the NTEWS.

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.

Career OneStop Certification Finder

EXECUTIVE SUMMARY

Dataset Name: **Certification Finder**

Publisher: U.S. Department of Labor, Employment and Training Administration (ETA)

Website: https://www.careeronestop.org/Toolkit/Training/find-certifications.aspx

Unit of Analysis: Credential

Quality Metrics

Metric	Rating	Summary Explanation
Non-response	Excellent	11,543 certifications covered in dataset, which is on the high end of estimates of the number of certifications available in the U.S. No known missing cases.
Coverage	Good	Coverage rates range from 24% to 99% for individual variables.
Granularity	Fair	Contains detailed O*Net and NAICS codes, but some variables (e.g., recertification requirements) lack details.
Consistency	Excellent	Other datasets ask similar questions and use similar reporting methodologies.
Timeliness	Excellent	There is no known delay between data being reported to ETA and publication to the website.
Integrity	Excellent	We did not identify any risks to data integrity.
Accessibility	Excellent	The data can easily be searched and browsed online, and anyone can download and work with the microdata immediately.
Interoperability	Good	Data can be matched to other sources on the name of the certification, though no widely accepted, unique identifier for certification organizations exists.
Suitability for Longitudinal Research	Below Average	One would need to have access to archived files (not published online) to see changes in the characteristics of certifications over time.
Overall Recommendation	Good	This is a high-quality source of information on one particular type of non-degree credential (industry certification).

Relevance to Use Cases

Use Case	Rating	Summary Explanation
Analyze the Overall Prevalence of NDCs	Fair	While Certification Finder helps us count the total number of certifications and certification organizations, it does not tell us about the number of individuals who earn or hold each certification.
Identify Which NDCs are Associated with Highest Earnings	Fair	Certification Finder does not contain information on the characteristics of individuals who attain certifications. However, it may be possible to link data on the average earnings in associated occupations to Certification Finder data to say something about the overall distribution of certifications in the labor market.
Identify Patterns of Inequality in NDC Attainment	Poor	Certification Finder does not contain information on the characteristics of individuals who attain certifications.

Use Case	Rating	Summary Explanation
Enrichment of NTEWS Microdata	Good	With some cleaning, data on certifications could be matched on to NTEWS data on individual certifications held by members of the skilled technical workforce to learn more about the quality and characteristics of their certifications.

DATA QUALITY ASSESSMENT – CERTIFICATION FINDER

Rubric for Assessing Dataset Quality

Each dataset subject to our data quality assessment will be evaluated according to a rubric as defined below. The answers to these questions will be used to create and publish profiles of each dataset, which will be available on the website of the GW Program on Skills, Credentials, and Workforce Policy.

1. Relevance

- What is the total number of items relevant to non-degree credentials?
  47 variables in use, including certification name, acronym, type, training/education/exam/renewal requirements, demand/accreditation status, NAICS and O*NET codes, and contact information of the certifying organization.

- What are the measures of NDC attainment like?
  The data developer defines a certification as "an award you earn to show that you have specific skills or knowledge in an occupation, industry, or technology". The following variables are relevant to measures of NDC attainment.
  - CERT_TYPE (certification type): core, advanced, skill, specialty or product/equipment (see here for definitions)
  - Requirements:
    - TRAINING: Is significant education or training is needed? Yes, No, or Unknown
    - EXPERIENCE: Is significant work experience is needed? Yes, No, or Unknown
    - EXAM: Is an exam required? Yes, No, or Unknown
    - RENEWAL: How many years until certification must be renewed?
    - CEU: Can it be renewed with Continuing Education Units? Yes, No, or Unknown
    - REEXAM: Can it be renewed with re-examination? Yes, No, or Unknown
    - CPD: Can it be renewed with Continuing Professional Development? Yes, No, or Unknown
    - CERT_ANY: Can it be renewed in multiple ways. Yes, No, or Unknown
  - ACCRED_ID (demand/accreditation status)
    - High demand: frequently mentioned in online job postings
    - Industry endorsed: endorsed by a major industry association that is not itself the developer of the certification
    - Related to the Job Corps training program
    - Identified in military Credentialing Opportunities On-Line (COOL) sites
    - Accredited by the American National Standards Institute (ANSI)
    - Accredited by the National Commission for Certifying Agencies (NCCA)
  - NAICSCODE and ONETCODE

- 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 Certification Finder is intended to help people identify professional certifications that may be useful for current or future employees.

a. Measuring the rate of attainment of non-degree credentials within the U.S. skilled technical workforce
   Not relevant. No data on the number of certified individuals
b. Measuring aggregate returns to non-degree credentials by credential type
   Not relevant. No data on returns of a certification. Not linkable to individual income information b/c no data on certified individuals.
c. Identifying disparities by race and gender in the attainment of non-degree credentials
   Not relevant. No data on the demographics of certified individuals.
d. Identifying which non-degree credentials are associated with the strongest labor market returns for individuals in the skilled technical workforce
   Relevant. Data available on demand/accreditation status as well as NAICS/O*NET codes.
e. Evaluating the effectiveness of public policies that support the attainment of non-degree credentials?
   Relevant. Data available on a certification's relation to Job Corps (a public training program), COOL (a military credentialing assistance programs), and two non-profit accreditation agencies, ANSI and NCCA.
f. *Other examples we might add?*

2. Non-response

- What is the frame of reference for the dataset, what population does the dataset attempt to cover?
  The data developer does not explicitly state the frame of reference/population of the dataset, but it seems that the dataset attempts to cover all professional certifications in the United States.

- What is the number of cases, and how does that number compare to known estimates of the relevant population?
  11,543 certifications in the dataset (before deduplication, as of 03/02/2022). We do not have an estimate for the total number of professional certifications in the United States.

- How does the publisher of the data ensure that data is collected for cases that should be in the dataset?
  We did not find information about the publisher's data collection practices.

- Do cases that we believe should exist in the microdata actually exist in the data?
  We do not see any sign of missing cases in our assessment of the data.

Summary assessment: Did the data publisher make adequate efforts to avoid non-response? Describe any efforts to avoid non-response and any evidence that non-response has been minimized.

3. Coverage

- What does the organization that creates or maintains this dataset do to minimize missing data?
  We did not see evidence of a specific strategy to minimize missing data.

- What percent of cases lack data for key variables of interest? (Direct assessment if not in metadata.)

Coverage rate for key variables ranges from 24% to 99%. See below for details. (11,543 records as of 03/02/2022)

Variable	Obs.	Coverage
TRAINING	8,597	74%
EXPERIENCE	8,415	73%
EXAM	11,450	99%
RENEWAL	6313	55%
CEU	6,040	52%
REEXAM	5,685	49%
CPD	5,294	46%
CERT_ANY	6,536	57%
CERT_TYPE	9,526	83%
ACCRED_ID	2,743	24%
NAICSCODE	8,921	77%
ONETCODE	8,748	76%

- What percent of the population of interest is in the dataset (if the size of the overall population is known or can be estimated)?
  We do not have an estimate for the population of professional certifications in the United States.

4. 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?
  There are five types of credentials, seven types of indicators of demand/accreditation status, seven trinary (yes/no/unknown) variables for training/education/exam/renewal requirements, and a continuous variable for the maximum years between two renewals. The NAICS code is at the six-digit level and the O*NET code is at the eight-digit level.

Summary assessment: Is this data granular enough (yes or no) to perform analyses for each of the use cases identified under "relevance"? How do we rate the overall granularity of the data (high, medium, low)?

The data is granular enough for analyses related to industry and occupations. It is not granular enough for analyses related to training/education/exam/renewal requirements as most variables of interest are only trinary (yes/no/unknown).

5. Timeliness

- How often is the dataset updated?
  According to the Data Source page on CareerOneStop, the dataset is updated on a rolling basis, though the technical document of the dataset notes that data is updated biannually. The most recent downloadable

version of the dataset is dated 03022022, and a 09032021 version, previously downloaded, is also available, so it seems that the update schedule is now on a rolling basis.

- Is data collected continuously or in waves? If in waves, what is the duration of those waves? Are some variables more frequently updated than others?
  Continuously. No information on whether some variables are more frequently updated than others.

- What is the time lag between when an event occurs, when it is recorded, and when that data is available to researchers?
  How soon a certification will be added to the dataset after its creation is unknown, but the dates when a certification is added to the dataset and when it is last updated are both available.

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

The dataset is updated on a rolling basis. We do not know how soon a certification will be added to the dataset after its creation. Researchers can assess the recency of a record through the dates when it is added and when it is last updated.

6. Integrity

- What are the risks to the integrity of this dataset?
  CareerOneStop, the data developer, is sponsored by the Employment and Training Administration (ETA) of the U.S. Department of Labor. We are not able to identify any special ties between a specific certifying organization/industry group and CareerOneStop/ETA.

- How are data outliers handled? (May be available from published documentation if not metadata.)
  It seems that data is recorded as-is. Most key variables are categorical. The only key variable that is continuous is the maximum years between two renewals, which ranges from 0 to 10. We do not know if the data developer tries to handle outliers in a certification's frequency of being mentioned in online job postings. This data is used to determine whether a certification is in high demand but is not disclosed in the dataset.

- 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.

7. Accessibility

- How do researchers access this dataset?
  The whole dataset is available for download. Researchers can also download part of the dataset by browsing specific occupations/industries or using the keyword search feature on the website.

- Are any variables or cases withheld from researchers? If so, does that withholding or censoring affect researchers' ability to use the data?
  Yes. SUPPRESS indicates whether a record contains confidential data that must be suppressed for public use. However, the type of suppressed information is unknown. DELETED indicates whether a record is deleted from the website but still exists in the database.

- Are there direct costs or indirect costs (e.g., training, resources) associated with accessing and using the data?
  Downloading the dataset is free. Currently the folder size is 38MB and the dataset can be opened with Microsoft Access or SQL.

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?

Yes, researchers can download the whole dataset for free and open it easily on most devices equipped with Microsoft Access or SQL.

8. Interoperability

- Is there a unique identifier for individual cases? If so, is it one that can be found in other datasets?
  Yes, it is CERT_ID (numerical). The variable is created exclusively for this dataset, but researchers may link records to other datasets through CERT_NAME (text), which is not unique to all records and indicates the existence of duplicates.

- 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).
  Certification names should be common and largely consistent among datasets of certifications.

- 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, both O*NET and NAICS 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).

Linkage on individual cases is possible through CERT_NAME, though deduplicating records and comparing them with certification names in other datasets can be demanding. Linkage to occupation and industry coding schemes is straightforward as both O*NET and NAICS codes are available. Records can also be linked to job training and accreditation programs like Job Corps, the military COOL, ANSI, and NCCA through corresponding indicators. The name and contact information (phone, email, address) of the certifying organization is also provided so linkage to certifying organizations is also possible.

9. Suitability for Longitudinal Research?

- Is the metadata consistent over time, at least for key variables?
  Yes for key variables, based on the two versions available (03022022 and 09032021). According to the technical document, three variables (EITHER, NSSB_URL, and CERT_URL) are not in use anymore.

- 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?
  For each downloadable version, the data is cross-sectional, and only the most recent version is open to download from the website. CERT_ID, consistent over time, can be used to link records of the same certification in different versions.

- How far back do administrative records from this dataset go?
  2002, according to the technical document.

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.

Data for Certification Finder has been collected for 20 years and key variables seems to be consistent over time. Only the most recent version is open to download from the website, adding to the difficulty of doing longitudinal research with this dataset. Researchers can link records of the same certification in different versions using CERT_ID, unique to each record and consistent over time.

Career OneStop License Finder

EXECUTIVE SUMMARY

Dataset Name: **License Finder**

Publisher: U.S. Department of Labor, Employment and Training Administration (ETA)

Website: https://www.careeronestop.org/Toolkit/Training/find-certifications.aspx

Unit of Analysis: Credential

Dataset Purpose: To create a comprehensive directory of licenses available to American workers to help job seekers and career changers identify potential new career pathways and learn how to prepare for new careers.

Quality Metrics

Metric	Rating	Summary Explanation
Non-response	Excellent	10,715 records. No known missing cases.
Coverage	Excellent	Coverage rates range from 86% to 100% for individual variables.
Granularity	Good	Contains detailed O*Net and NAICS codes, but some variables (e.g., renewal requirements) lack details.
Consistency	Excellent	Other datasets (e.g., the DOL Certification Finder) ask similar questions and use similar reporting methodologies.
Timeliness	Excellent	There is no known delay between data being reported to ETA and publication to the website.
Integrity	Excellent	We did not identify any risks to data integrity.
Accessibility	Excellent	The data can easily be searched and browsed online, and anyone can download and work with the microdata immediately.
Interoperability	Good	Data can be matched to other sources on the name of the license, though no widely accepted, unique identifier for state licensure bodies exists.
Suitability for Longitudinal Research	Below Average	One would need to have access to archived files (not published online) to see changes in the characteristics of licenses over time.
Overall Recommendation	Good	This is a high-quality source of information on one particular type of non-degree credential (occupational licenses).

Relevance to Use Cases

Use Case	Rating	Summary Explanation
Analyze the Overall Prevalence of NDCs	Fair	While License Finder tells us about the universe of licenses that exist, it does not tell us about the number of individuals who earn or hold each certification.
Identify Which NDCs are Associated with Highest Earnings	Fair	License Finder does not contain information on the characteristics of individuals who attain licenses. However, it may be possible to link data on the average earnings in associated occupations to License Finder data to say something about the overall distribution of licenses in the labor market.
Identify Patterns of Inequality in NDC Attainment	Poor	License Finder does not contain information on the characteristics of individuals who attain licenses.

Use Case	Rating	Summary Explanation
Enrichment of NTEWS Microdata	Good	With some cleaning, data on licenses could be matched on to NTEWS data on individual licenses held by members of the skilled technical workforce to learn more about the quality and characteristics of their licenses.

DATA QUALITY ASSESSMENT

1. Relevance

- What is the total number of items relevant to non-degree credentials?
  32 variables for each license record. Information includes license titles, types, descriptions, states/territories, licensing agencies, application requirements, active, status, and NAICS and O*NET codes. The dataset also includes state FIPS, active status, and O*NET codes for 14 license compacts.

- What are the measures of NDC attainment like?
  The definition of an occupational license as used in the Current Population Survey and adopted by License Finder is that a license
  - Is a credential awarded by a governmental licensing agency based on pre-determined criteria
  - The criteria may include some combination of degree attainment, certifications, educational certificates, assessments (including state-administered exams), apprenticeship programs, or work experience
  - Conveys a legal authority to work in an occupation

  The following variables are relevant to measures of NDC attainment.
  - License type: Stand-alone license/Registry/Tied to business/Secondary license (other license is prerequisite)/Preliminary or temporary license/Undetermined
  - NAICS and O*NET codes
  - Application requirements:
    - Exams: No exam/State exam required/Third-party exam required/Both state and third-party exams required/Choice of state or third-party exam/Undetermined
    - Education: No education required/Specific course required/Degree required/Undetermined
    - Continuing education: No CE requirement/CE required/Undetermined
    - Certification: No mention of certification/Certification may substitute for license requirements/Certification required/Undetermined
    - Experience: No experience/Affidavit or referral/Experience/Current employment/Undetermined
    - Criminal record: No criminal record requirements/Specific type of conviction prohibited/Felony convictions prohibited/Any conviction prohibited/Background check required/Undetermined
    - Physical: No physical requirements/Vision test required/Physical exam/More significant physical requirements/Undetermined
    - Veteran preference: No veteran preference/Undetermined/A temporary license available to military and spouses until formal license approval/Advisers or additional guidance is available for military and spouses/Licensure by endorsement is available for military and spouses/Expedited processing is available for military and spouses/Expedited processing is available for military and spouses, and a separate licensure by endorsement process

occurs/Expedited processing is available for military and spouses, and no background check is required/Expedited processing is available for military and spouses, with a temporary license available in the interim/Fees are reduced for military and spouses/Fees are reduced and a temporary license available in the interim/Fees are reduced and expedited processing is available for military and spouses/Fees are reduced and expedited processing is available for military and spouses, with a temporary license available in the interim/Military and spouses are exempt from licensure requirements

- Active status: Active/No new licenses issued/Replaced/No longer licensed/Undetermined

- Are there any indicators related to NDC attainment that are unique to this dataset?
  Yes. It has a unique classification of license types and provides rich information on application requirements.

- Are there indicators of other phenomena that could be of sociological significance?
  No. The License Finder only contains data on licenses and the organizations that issue them.

- 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 License Finder is intended to help users find information about occupational licenses that states require for some jobs.
  a. Measuring the rate of attainment of non-degree credentials within the U.S. skilled technical workforce
     Not relevant. No data on the number of licensees.
  b. Measuring aggregate returns to non-degree credentials by credential type
     Not relevant. No data on returns of a license. Not linkable to individual income information b/c no data on licensees.
  c. Identifying disparities by race and gender in the attainment of non-degree credentials
     Not relevant. No data on the demographics of licensees.
  d. Identifying which non-degree credentials are associated with the strongest labor market returns for individuals in the skilled technical workforce
     Relevant. NAICS and O*NET codes are available.
  e. Evaluating the effectiveness of public policies that support the attainment of non-degree credentials?
     Relevant. The dataset provides information on important regulatory issues such as the type, application requirements, and active status of a license (or a license compact).
  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 dataset attempts to cover all occupational licenses in the United States.

- What is the number of cases, and how does that number compare to known estimates of the relevant population?
  10,715 license records in the dataset (before deduplication, as of 03/2022). We do not have an estimate for the total number of occupational licenses in the United States.

- How does the publisher of the data ensure that data is collected for cases that should be in the dataset?
  Analyst Resource Center (ARC), the data developer, collects occupational license data from each U.S. state and territory and combine it with additional data from federal agencies and web-scraping. The data goes through a central clean-up process that standardizes occupational coding, adds likely licenses, and ensures consistent structure.

  ARC uses text parsing and information from the Center for State Occupational Regulation (CSOR), License2Work, the National Center for State Legislatures (NCSL), and the Military Spouse Portability Examination Report from UMN to obtain data for active status and application requirements. Data related to license compacts are obtained via internet search and is updated on an as-needed basis. Data related to industry is identified based on license description and details. When no industry is specified, it is assumed that the occupation requires a license across all industries.

  ARC accepts requests for changes or additions to database through email (arc.deed@state.mn.us) or telephone (651-259-7398).

- Do cases that we believe should exist in the microdata actually exist in the data?
  ARC acknowledges that some states do not participate in the data collection effort and there is no way to guarantee that each state collects all of the license information for their state. In addition, inaccurate reporting also contributes to the missing data problem.

- What percent of cases lack data for key variables of interest? (Direct assessment if not in metadata.) In the flat file (explained in "Accessibility"), the missing rate is 0.18% for O*NET codes and 14% for license types, but the missing rate for license types is zero in the relational file. For all other key variables, the missing rate is zero (as of 03/2022).

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.

Though a full coverage of all occupational licenses in the U.S. cannot be guaranteed, the data developer has made adequate effort to ensure that data missed from state submissions are filled with information from various sources. The missing rate for key variables is substantially zero, and the dataset is sufficient for each use case identified as relevant in "Relevance".

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?
  Variables for license types, application requirements, and active status, as listed in Section 1, have a high level of granularity. Both two-digit and six-digit NAICS codes are available, and the O*NET code is at the eight-digit level.

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

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

4. Timeliness

- How often is the data set updated?
  Officially every four to six months.

- 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?
  License information is collected from each state by the Analyst Resource Center and available for download on CareerOneStop. States are expected to submit revisions every two years, and new information is updated on CareerOneStop every four to six months, typically in March, July, and November, as new information is received. Some records may be more frequently updated than others as the update frequency varies with states and occupational licenses.

- What is the time lag between when an event occurs, when it is recorded, and when that data is available to researchers?
  The maximum time lag between when a state submits new information and when the information is updated in the dataset seems to be six months.

  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 four to six months. The maximum time lag between data collection and data availability seems to be six months.

5. Integrity

- What are the risks to the integrity of this dataset?
  Most of the data is reported by state agencies, which may have different levels of motivation in reporting data accurately and timely.

- How are data outliers handled? (May be available from published documentation if not metadata.) It seems that data is recorded as-is. Outliers are not a significant issue in this dataset as all key variables are categorical.

- 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

xxx

- How do researchers access this dataset?
  The whole dataset is available for download in Microsoft Access in two different file types, relational and flat. The flat file contains all available information in one table, but because many licenses are coded to multiple O*NET codes, there are duplicates.

- Are any variables or cases withheld from researchers? If so, does that withholding or censoring affect researchers' ability to use the data?
  The dataset contains a table of the number of licenses awarded for a selected occupation in each state, but currently there is no records in that table in the published version of the dataset.

- Are there direct costs or indirect costs (e.g., training, resources) associated with accessing and using the data?
  Downloading the dataset is free. Currently the file size is 26MB for the relational file and 137MB for the flat file. The data developer recommends using the dataset in Microsoft Access because when exporting data in other formats, such as Microsoft Excel, the length of the description fields may cause formatting issues.

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?

Yes, researchers can download the whole dataset for free and open it easily on most devices equipped with Microsoft Access.

7. Interoperability

- Is there a unique identifier for individual cases? If so, is it one that can be found in other datasets? No, the license ID in this dataset is not a unique identifier, nor is the combination of license ID and state FIPS. Inaccurate reporting of state agencies adds to the difficulty of data cleaning and is a source of duplicates in the dataset.

- 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 found.

- 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, both O*NET and NAICS 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).

Linkage on individual cases is poor as there is no unique identifier, but still possible through license names. Linkage to occupation and industry coding schemes is straightforward as both O*NET and NAICS codes are available.

8. Suitability for Longitudinal Research?

- Is the metadata consistent over time, at least for key variables?
  Yes for key variables, based on the two versions available (03/2022 and 11/2021).

- 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?
  For each downloadable version, the data is cross-sectional, and only the most recent version is open to download from the website. Linkage among the records for same license over time is difficult as no unique identifier exists.

- How far back do administrative records from this dataset go?
  According to the dataset's technical document, data collection began in 1997. However, in a 2019 quality review, the reviewer notes that available prior versions date back to 2018. The reviewers acknowledged the complexity of collecting longitudinal data but also expressed interest in creating a time series.

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.

Data for License Finder has been collected for 25 years and key variables seems to be consistent over time. Only the most recent version is open to download from the website and no unique identifier is available, adding to the difficulty of linking records from different time periods and doing longitudinal research with this dataset.

Army COOL (Credentialing Opportunities Online)

EXECUTIVE SUMMARY

Key Dataset Facts

Dataset Name: **Army COOL (Credentialing Opportunities Online)**

Publisher: U.S. Department of Defense

Website: https://www.cool.osd.mil/army/

Unit of Analysis: Credential

Quality Metrics

Metric	Rating	Summary Explanation
Non-response	Excellent	Excellent All credential entries appear to be complete.
Coverage	N/A	Since COOL attempts to identify all certifications and closely related credentials (generally NOT certificates) relevant to military work, the sampling frame is inherently subjective.
Granularity	Good	Contains ample detail about the characteristics of each credential, including qualitative/narrative descriptions of available preparation materials, prerequisites, and so on.
Consistency	Excellent	Similar to DOL Certification Finder entries, but more complete for certain credentials. However, the population of credentials is not identical to DOL's since COOL only attempts to cover those credentials deemed relevant to military work and skillsets.
Timeliness	Excellent	There is no known delay between data being reported to DOD and publication to the website.
Integrity	Excellent	Excellent We did not identify any risks to data integrity.
Accessibility	Fair	The data can easily be searched and browsed online, however microdata would have to be scraped from the website or obtained by contacting DOD.
Interoperability	Good	Data can be matched to other sources on the name of the certification, though no widely accepted, unique identifier for certification organizations exists.
Suitability for Longitudinal Research	Poor	One would need to have access to archived files (not published online) to see changes in the characteristics of certifications over time.
Overall Recommendation	Fair	This is a high-quality source of information on industry certifications that are covered, though the coverage is of a limited portion of the labor market.

Relevance to Use Cases

Use Case	Rating	Summary Explanation
Analyze the Overall Prevalence of NDCs	Poor	Since COOL only covers occupations relevant to the military, it represents an inherently incomplete sample of the overall population.
Identify Which NDCs are Associated with Highest Earnings	Poor	COOL does not contain data on earnings associated with particular credentials.
Identify Patterns of Inequality in NDC Attainment	Poor	COOL does not contain information on the characteristics of individuals who attain certifications.
Enrichment of NTEWS Microdata	Fair	With some cleaning, data on certifications in COOL could be paired with NTEWS microdata to learn more about the credentials held by, or available to, service members, veterans, and others with applicable skillsets.

DATA QUALITY ASSESSMENT

Rubric for Assessing Dataset Quality

1. Relevance

- What is the total number of items relevant to credentials?
  For each credential, DOD COOL contains information on its title, type, credentialing agency, indicator for in-demand credentials based on CareerOneStop's analysis, indicator for GI bill applicability, accreditation status, related military personnel/services/occupations, related civilian occupations, level of relevance to each military occupation, attainability, eligibility requirements (experience/education/fee), exam requirements and administration, and recertification requirements. Different versions of COOL exist for each service branch; this assessment focuses on the Army's version of COOL.
- What are the measures of credential attainment like?
  Civilian occupational licenses, certifications, and apprenticeships.
- Are there any indicators related to education attainment that are unique to this dataset?
  GI bill indicator, related military personnel/services/occupations, and level of relevance to each military occupation for each credential.
- Are there indicators of other phenomena that could be of sociological significance? No.
- 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.)
  COOL is an information resource for military service members to learn what civilian credentials pertain to their military training and experience and resources available to help them attain the credentials, and for civilian personnel to learn about professional development opportunities available in their career areas. It also provides tools that can be used to research military occupations and related credentials.
  a. Measuring the rate of attainment of credentials within the U.S. skilled technical workforce Not relevant. The dataset does not include the number of credential holders.
  b. Measuring aggregate returns to credentials by credential type

Not relevant. The dataset does not include employment or earnings information.

c. Identifying disparities by race and gender in the attainment of credentials Not relevant. The dataset does not include demographic information.
d. Identifying which credentials are associated with the strongest labor market returns for individuals in the skilled technical workforce
Somewhat relevant. The dataset links to CareerOneStop's analysis of whether the credential is frequently mentioned in online job postings.
e. Evaluating the effectiveness of public policies that support the attainment of credentials?
Somewhat relevant. Policymakers can use the dataset to assess which credentials are most relevant to professionalizing military service members and enhancing their ability to transition to the civilian workforce upon completion of military service.
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?
COOL attempts to cover all civilian occupational licenses, certifications, and apprenticeships that meet DOD's Credentialing Standards and are relevant to military service members in the United States.
- What is the number of cases, and how does that number compare to known estimates of the relevant population?
As of September 2022, DOD COOL includes 2,396 entries for related civilian credentials.
- How does the publisher of the data ensure that data is collected for cases that should be in the dataset?
Detailed information about DOD's data collection procedure is not published. DOD has relied on a contractor, Solutions for Information Design (SOLID), to assist with this work. We have no reason to doubt that diligent efforts have been made to identify all credentials of value to service members.
- Do cases that we believe should exist in the microdata actually exist in the data?
No. The "universe" of credentials that should be in COOL is naturally limited given that military careers will inherently not be applicable to every possible civilian career and credential pathway (even if it is a larger universe than one would find in just about any other employer).
- What percent of cases lack data for key variables of interest? (Direct assessment if not in metadata.) All credential entries appear to be "filled in" completely, so 100%.

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?
    - Accreditation: National Accreditation Board (ANAB) of the American National Standards Institute (ANSI)/National Commission for Certifying Agencies (NCCA)/International Certification

Accreditation Council (ICAC)/Accreditation Board for Specialty Nursing Certification (ABSNC)/International Accreditation Service (IAS)
  - Credential type: National Certification/Federal License/State License
  - Related personnel: Enlisted/Officer/Enlisted and Officer
  - Related services: Army/Navy/Air Force/Coast Guard/Marine Corps/DoD Civilian
  - Relevance level: Most/Some/Other
  - Attainability: High/Medium/Low
- 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?
  We assume that the dataset is updated continuously; however, there is no published schedule or timeframe for updates.
- 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?
  Continuous
- What is the time lag between when an event occurs, when it is recorded, and when that data is available to researchers?
  Unknown

5. Integrity

- What are the risks to the integrity of this dataset?
  Credentialing bodies can submit information to COOL to have their credentials included. In theory, credentialing bodies have an incentive to misreport information in their favor, though the COOL team conducts an information review process.
- How are data outliers handled? (May be available from published documentation if not metadata.) Most of the data in COOL is not continuous in nature so there are few if any true "outliers," however we are not aware of any efforts to suppress credentials with very high or low values.
- 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 can identify.

6. Accessibility

- How do researchers access this dataset?
  Data can be accessed by search but cannot be downloaded.
- Are any variables or cases withheld from researchers? If so, does that withholding or censoring affect researchers' ability to use the data?

No. COOL is intended to be used by the public. There is no difference between the version accessible to DOD and the version that can be accessed anywhere in the world. In fact, DOD encourages public use to help prospective service members see the applicability of military training to civilian careers.

- Are there direct costs or indirect costs (e.g., training, resources) associated with accessing and using the data?
  Searching in the database is free. DOD does not publish a master microdata file in a research-friendly format, which may be an obstacle to some research uses - though presumably it could be possible to obtain such a file from DOD as the information is unclassified.

7. Interoperability

- Is there a unique identifier for individual cases? If so, is it one that can be found in other datasets? Not in the published webpages.
- 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?
  MOC codes and titles for military occupations and O*Net titles for relevant civilian occupations are available.

8. Suitability for Longitudinal Research?

- Is the metadata consistent over time, at least for key variables?
  Apparently so. However, it is not possible to easily view past versions/editions of COOL barring the use of archived versions of the website (as may exist in the Internet Archive's "Wayback Machine").
- 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. The COOL team updates information about credentials and remove credentials that are no longer relevant. However, published webpages only show the latest information of credentials that remain relevant.
- How far back do administrative records from this dataset go?
  The COOL initiatives began in 2002 when the Army launched the first COOL web site. The Navy launched its COOL in 2006, and the Marine Corps and Air Force launched their own sites in 2014. In 2019, the Coast Guard and DOD Civilian sites completed the COOL team.

Eligible Training Provider Performance Results (ETPPR)

EXECUTIVE SUMMARY

Key Dataset Facts

Dataset Name: **TrainingProviderResults.gov Data Extract**

Publisher: Employment and Training Administration, U.S. Department of Labor

Website: www.trainingproviderresults.gov

Unit of Analysis: Credential Issuer

Purpose of Dataset: To bring transparency to the marketplace for training programs, especially those which are eligible for student financial assistance with tuition and related expenses through an Individual Training Account (ITA), by providing data reported by training institutions and collected by states on the characteristics of each program and information on the labor market experiences (earnings, employment status) of individuals who complete those programs in each of four quarters following credential completion.

Quality Metrics

Metric	Rating	Summary Explanation
Non-response	Good	The population of 75,676 training providers is a plausible estimate of all training providers and programs eligible for WIOA support in the United States
Coverage	Fair	Tremendous variation, from 3% (percent of graduates who earned a credential via WIOA) to 100% (training provider street address). Coverage tends to be weakest for items related to average earnings after completion of a credential.
Granularity	Excellent	Specific data on each credential, including relevant O*Net and CIP codes. Precise data on the location of each training provider, including latitude and longitude.
Consistency	N/A	We have not been able to find datasets that would lend themselves to an “apples to apples” comparison with TrainingProvdierResults.gov.
Timeliness	Good	As of 2021, the data is being refreshed on an annual basis. There is some variation between states in the lag time between data collection and publication.
Integrity	Excellent	Excellent We did not identify any risks to data integrity.
Accessibility	Excellent	Anyone can download the dataset from the Department of Labor without cost.

Metric	Rating	Summary Explanation
Interoperability	Excellent	One can probably link to IPEDS on the basis of the institution name and credential name, though there is no numerical direct identifier of training providers that would facilitate broader linkages across relevant datasets such as the PIRL.
Suitability for Longitudinal Research	Fair	The dataset is new, published for the first time in 2020, but could have value for longitudinal research as time passes if prior years' files are properly archived by DOL.
Overall Recommendation	Good	This is a high-quality source of information on many different types of credentials. While it does not cover the entire universe of training providers in the United States, it does give valuable information about a large number of programs that could be useful for purposes beyond evaluating the experiences of the subset of students at these programs who receive WIOA support.

Relevance to Use Cases

Use Case	Rating	Summary Explanation
Analyze the Overall Prevalence of NDCs	Fair	As this is not a representative sample of all training providers and credential issuers, enrollment information reported here will be of limited value for developing population-level estimates.
Identify Which NDCs are Associated with Highest Earnings	Good	While limited to data on the earnings of individuals who were WIOA participants – a relatively disadvantaged subset of the overall student population at many institutions – it does provide high quality data in cases where data is reported.
Identify Patterns of Inequality in NDC Attainment	Fair	While containing data on labor market outcomes, those outcomes are not broken out by race, ethnicity, sex, or other demographic variables of interest.
Enrichment of NTEWS Microdata	Fair	The primary challenge to linking with NTEWS would be precisely identifying a training provider in both datasets. While NTEWS collects data on the names of individual credentials, data may need to be cleaned and there may be cases of ambiguously named credentials that cannot be positively identified.

DATA QUALITY ASSESSMENT

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
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

Variable	Missing	%Missing	Variable	Missing	%Missing
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).

National Labor Exchange Research Hub (NLx)

EXECUTIVE SUMMARY

Key Dataset Facts

Dataset Name: **NLx (National Labor Exchange) Research Hub**

Publisher: Direct Employers Association; National Association of State Workforce Agencies

Website: https://nlxresearchhub.org/

Unit of Analysis: Job Posting/Vacancy

Purpose of Dataset: To enable researchers to conduct analyses using "big data" on job postings aggregated from multiple job and career search websites and workforce boards/agencies. This data is intended to enable research that informs policy and/or guides the decision-making of job and credential seekers.

Quality Metrics

Metric	Rating	Summary Explanation
Non-response	Good	There are about 4 million job postings in the NLx at any given moment. It is intended to be comprehensive. However, the full "denominator" or number of vacancies that exist is difficult to ascertain.
Coverage	Fair	Employers vary greatly in the amount of information that they put in job descriptions, ranging from a few sentences to several paragraphs. There is rich and ample data, but it is not consistent.
Granularity	Excellent	Some job postings are quite specific about what they are looking for. Names and physical locations of employers exist in the dataset, though are not always provided to researchers. Data on the O*Net codes associated with job postings exists, but there are limitations on its use by researchers.
Consistency	N/A	We have not been able to find datasets that would lend themselves to an "apples to apples" comparison with the NLx.
Timeliness	Excellent	The data is refreshed at least daily.
Integrity	Excellent	Excellent We did not identify any risks to data integrity.
Accessibility	Excellent	There is an application procedure that may take several hours to complete. Access to the microdata is through an API; while sample Python code is provided to approved researchers to access the data, significant technical expertise is required.
Interoperability	Good	There is certainly potential to establish linkages at various levels of geography, occupation, and employer name once the data has been suitably cleaned or processed. However, the data would require cleaning and special approval from the data owners may be required for some linkages.
Suitability for Longitudinal Research	Excellent	Data goes back several years and is updated on a daily basis enabling one to correlate job postings with the business cycle.

Metric	Rating	Summary Explanation
Overall Recommendation	Excellent	This is a unique source of data that, with sufficient investment in time and effort, could help researchers understand they dynamics of the relationship between workers and employers.

Relevance to Use Cases

Use Case	Rating	Summary Explanation
Analyze the Overall Prevalence of NDCs	Good	While this dataset tells us nothing about who holds credentials, it can tell us a lot about who is requesting credentials – which may be a suitable proxy in some situations for credential prevalence.
Identify Which NDCs are Associated with Highest Earnings	Good	Some but not all job postings have information about the potential salary (or range of salaries) that candidates may be considered for.
Identify Patterns of Inequality in NDC Attainment	Poor	Since the unfilled job posting is the unit of analysis, we do not know whether the jobs listed are going to members of particular demographic groups.
Enrichment of NTEWS Microdata	Good	There are applications in terms of defining the universe of credentials being requested by employers, the value employers are associating with particular credentials, and the availability of jobs in particular occupations and locations, all of which can be paired at various levels of aggregation with NTEWS microdata.

DATA QUALITY ASSESSMENT

Rubric for Assessing Dataset Quality

1. Relevance

- What is the total number of items relevant to credentials?
  There are 51 variables in the dataset, those most relevant to NDC attainment include job ID, job title, job description, job URL, O*Net code, NAICS code, number of positions available, job schedule, job shift, expected number of hours per week, salary unit, salary minimum and maximum, minimum education required, minimum experience required, license requirements, training requirements, application methods, job posting location, date and time the job was first acquired by DirectEmployers, date of most recent update, date of expiration; company name, contact, location Federal Contractor status, and Federal Employer Identification Number.
- What are the measures of credential attainment like? Degree, license, and training programs.
- Are there any indicators related to education attainment that are unique to this dataset? Job salary and hours information.
- Are there indicators of other phenomena that could be of sociological significance?
  There is potential to draw out a huge amount of data from the job posting descriptions. One could

theoretically mine data on certifications and other credentials required, skill requirements, experience requirements, benefits, and company policies, among other factors.

- 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.)
  A partnership between the National Association of State Workforce Agencies (NASWA) and DirectEmployers, NLx Research Hub is a cloud-hosted warehouse of jobs data sourced from the National Labor Exchange. NLx Research Hub is created to increase the amount of labor market information in the U.S. to facilitate the recruitment, hiring, and training opportunities of U.S. workers and deepen partnerships between industry, government, and academia by enhancing the infrastructure to support the convergence of research, education, and talent pipelines.
  a. Measuring the rate of attainment of credentials within the U.S. skilled technical workforce Not relevant. The dataset does not include individual or aggregate level information of employees or jobseekers.
  b. Measuring aggregate returns to credentials by credential type
     Somewhat relevant. Researchers can combine credential requirement and salary information to assess the aggregate returns to a credential.
  c. Identifying disparities by race and gender in the attainment of credentials Not relevant. The dataset does not include demographic information.
  d. Identifying which credentials are associated with the strongest labor market returns for individuals in the skilled technical workforce
     Somewhat relevant. Researchers can combine job posting counts, credential requirement and salary information to assess the employability and expected salary a credential.
  e. Evaluating the effectiveness of public policies that support the attainment of credentials?
     Somewhat relevant. Researchers can use the dataset to inform policymakers of the most needed credentials.

2. Coverage

- What is the frame of reference for the dataset, what population does the dataset attempt to cover? NLx attempts to cover all real job postings in the United States.

- What is the number of cases, and how does that number compare to known estimates of the relevant population?
  The database includes approximately 300,000 employers, 4 million daily job postings, and 75 million historical job postings, including openings from small- and medium-sized employers.

- How does the publisher of the data ensure that data is collected for cases that should be in the dataset?
  All 50 state workforce agencies, plus District of Columbia, Guam, Puerto Rico, and the U.S. Virgin Islands participate in NLx. NLx uses indexing (a.k.a. scraping or spidering) to extract job postings from the career sites of over 18,000 employers. The indexed employer community includes both DirectEmployers member companies and nonmembers who would like their jobs to appear in the NLx. State workforce agencies can help increase the total number of NLx job openings by identifying indexable corporate sites and notifying the NLx operations team.

  NLx gathers currently available and unduplicated job opportunities and takes validation seriously to ensure

only real jobs from verified employers are made available. Jobs with requirements to purchase training or equipment to secure employment, unpaid positions, commission-only positions, multi-level marketing positions, franchise opportunities, and companies currently involved in a strike or labor dispute or of questionable ethics or illegal are not included.

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?
  - Job posting location: zip code, city, state, country
  - Company location: address, zip code, city, state, country
  - Job schedule: part-time/full-time/flexible
  - Shift: day shift/night shift/swing shift
  - O*Net code:
  - NAICS code
- Is this data granular enough (yes or no) to perform analyses for each of the use cases identified under "relevance"?
  Yes, though the data may need serious cleaning and/or organizing.

4. Timeliness

- How often is the dataset updated? Daily.
- 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?
  Continuously.
- What is the time lag between when an event occurs, when it is recorded, and when that data is available to researchers?
  The NLx feed is refreshed on a daily basis through a "kill and fill" process. When a job is taken off a corporate website, state job bank, or USAjobs.gov, it will no longer be made available for viewing on NLx. This usually happens within one day of the job being removed from the source site.

5. Integrity

- What are the risks to the integrity of this dataset?
  None that we are able to identify. NLx takes validation process to ensure that scams and illegal jobs are excluded from the database.
- How are data outliers handled? (May be available from published documentation if not metadata.) There is no effort to exclude 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 need to complete and submit a data trust request form to request data from NLx.

- Are any variables or cases withheld from researchers? If so, does that withholding or censoring affect researchers' ability to use the data?
  Company's name, Federal Employer Identification Number, address line 1 and O*Net code of the job posting are restricted in the dataset. This limits researcher's ability to link job postings to companies and occupations.

- Are there direct costs or indirect costs (e.g., training, resources) associated with accessing and using the data?
  Researchers need to fill in the request form, provide supporting documents, and wait for approval from the NLx. Denied request cannot be resubmitted unless modified to account for reasons for denial.

7. Interoperability

- Is there a unique identifier for individual cases? If so, is it one that can be found in other datasets?
  Yes. There are two unique job IDs, one utilized by the Data Warehouse and another assigned by DirectEmployers. Linkages to industry, occupation, and company are possible but can be limited by data restrictions described in Section 6.

- 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, NAICS and O*Net codes are available, though O*Net code is restricted.

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?
  It is continuous. There are no "waves." One can look back in time and download data extracts referring to specific time periods in specific places, though if analyzing the data offline there may be computational challenges.

- How far back do administrative records from this dataset go?
  Approximately 2010.

Post-Secondary Employment Outcomes (PSEO)

EXECUTIVE SUMMARY

Key Dataset Facts

Dataset Name: **PSEO (Post-Secondary Employment Outcomes)**

Publisher: U.S. Census Bureau

Website: https://lehd.ces.census.gov/data/pseo_experimental.html

Unit of Analysis: Credential

Purpose of Dataset: To publish data on employment and earnings outcomes associated with specific degrees and fields of study in participating state systems of higher education.

Quality Metrics

Metric	Rating	Summary Explanation
Non-response	Good	There are 234,951 records for graduate earnings in 17 states. This appears to represent a significant portion of the credential-field of study-institution-year groupings we would expect to see in the 17 participating states, keeping in mind that some private institutions are excluded.
Coverage	Fair	Coverage is modest for the critical variables of interest. We have 46% coverage for graduate earnings and 64% coverage for the occupation in which one is employed for the first year after degree completion.
Granularity	Good	Sufficiently granular at the institution level in terms of degree field, though some are truncated at the 2-digit CIP level.
Consistency	N/A	We have not been able to find datasets that would lend themselves to an "apples to apples" comparison with PSEO.
Timeliness	Good	While we have not been able to find an official update schedule, the most recent update was in October 2021.
Integrity	Excellent	Excellent We did not identify any risks to data integrity.
Accessibility	Excellent	Anyone can download the dataset from the Census Bureau free of charge.
Interoperability	Good	One can probably link to IPEDS on the basis of the institution name and field of study. Since the data is aggregated at the program-institution-year level, there is no option to link individual-level data.
Suitability for Longitudinal Research	Good	The dataset has been updated at least annually since 2018, and historical files can be downloaded from the Census website. Currently, very few institutions have associated data on earnings 5 and 10 years after graduation, but this is likely to grow with time as the dataset ages.

Metric	Rating	Summary Explanation
Overall Recommendation	Good	This is a high-quality, innovative dataset that is unique for its ability to map "flows" between fields of study and occupations. While limited to participating institutions, the quality of earnings data is rich where available.

Relevance to Use Cases

Use Case	Rating	Summary Explanation
Analyze the Overall Prevalence of NDCs	Fair	While one could obtain enrollment data on specific institutions from this dataset, the data does not tell us about non-degree credentials other than certificates and we are not getting a full picture of all certificates issued even by public institutions within a state. In addition to non-participating institutions, even non-credit programs at participating institutions are missing.
Identify Which NDCs are Associated with Highest Earnings	Excellent	Bearing in mind the limits of the PSEO population size and included credential types, the dataset is great for comparing earnings between credentials and fields of study.
Identify Patterns of Inequality in NDC Attainment	Fair	While containing data on labor market outcomes, those outcomes are not broken out by race, ethnicity, sex, or other demographic variables of interest.
Enrichment of NTEWS Microdata	Fair	It may be possible to link PSEO outcomes data to NTEWS responses to the question that asks respondents to identify institutions attended by name. This might unlock interesting opportunities to examine the characteristics of individuals who underperform or outperform their peers from the same degree or certificate program.

DATA QUALITY ASSESSMENT

Rubric for Assessing Dataset Quality

1. Relevance

- What is the total number of items relevant to non-degree credentials?
  There are two PSEO datasets: Graduate Earnings and Employment Flows. They share 12 identifiers including the aggregation level, institution, degree level and field, graduation cohort, census division, and industry. Graduate Earnings additionally contains 15 indicators including 25th/50th/75th percentile earnings, counts of employed graduates, and counts of graduates 1/5/10 year(s) after graduation. Employment Flows additionally contains 9 indicators including counts of employed graduates, counts of graduates employed in the educational institution's state, and counts of jobless or marginally employed graduates 1/5/10 year(s) after graduation.
- What are the measures of NDC attainment like?
  Degree level and degree field. See "Granularity" for details.
- Are there any indicators related to education attainment that are unique to this dataset?

Yes, for each graduation cohort of a program/institution, PSEO provides data on its earnings and employment one/five/ten year(s) after graduation.

- Are there indicators of other phenomena that could be of sociological significance? No.

- 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.)
  PSEO is intended to inform current and prospective students on the labor market returns of attending a program/institution.
  a. Measuring the rate of attainment of non-degree credentials within the U.S. skilled technical workforce
     Relevant. PSEO provides graduate count data from the Integrated Postsecondary Education Data System (IPEDS).
  b. Measuring aggregate returns to non-degree credentials by credential type
     Relevant. PSEO provides both earnings and employment data.
  c. Identifying disparities by race and gender in the attainment of non-degree credentials Not relevant. No data on race and gender of graduates.
  d. Identifying which non-degree credentials are associated with the strongest labor market returns for individuals in the skilled technical workforce
     Relevant. PSEO provides data on earnings, employment, and industry.
  e. Evaluating the effectiveness of public policies that support the attainment of non-degree credentials?
     Relevant. PSEO can be used to evaluate labor market returns of a certificate that is being/can be supported by public policy.

2. Coverage

- What is the frame of reference for the dataset, what population does the dataset attempt to cover?
  PSEO attempts to cover all persons who received a degree or certificate from an institution.

- What is the number of cases, and how does that number compare to known estimates of the relevant population?
  As of October, PSEO covers 17 states, with 234,951 records in Graduate Earnings and 16,953,720 records in Employment Flows. See this map for graduate coverage rate by state.

- How does the publisher of the data ensure that data is collected for cases that should be in the dataset?
  PSEO is created by merging graduation files submitted by educational institutions with administrative data on jobs collected by the Longitudinal Employer-Household Dynamics Program (LEHD).

  Post-graduate population coverage: PSEO includes only graduates of in-scope institutions. Students who enroll but do not graduate are omitted from the statistics. Some graduates are also omitted from the earnings and employment outcome statistics because of insufficient labor market attachment in the reference year.

  Employment coverage: LEHD data covers over 96% of employment in the United States. These job records are supplemented with Census Bureau surveys and other federal agency administrative records to supply additional information on the characteristics of the workers and firms.

- Do cases that we believe should exist in the microdata actually exist in the data? According to the technical documentation, less than 1% of graduates are omitted from the published statistics due to poor quality of the personal identifier.

- What percent of cases lack data for key variables of interest? (Direct assessment if not in metadata.) All identifiers have no missing value. For earnings and employment variables, coverage is low to modest and decreases with the number of years after graduation. The coverage is extremely low for the count of jobless or marginally employed graduates. See below for details.

Graduate Earnings		Employment Flows	
Variable	**Coverage**	**Variable**	**Coverage**
y1_p25_earnings	46%	y1_grads_emp	64%
y1_p50_earnings	46%	y1_grads_emp_instate	64%
y1_p75_earnings	46%	y5_grads_emp	42%
y1_grads_earn	46%	y5_grads_emp_instate	42%
y5_p25_earnings	31%	y10_grads_emp	27%
y5_p50_earnings	31%	y10_grads_emp_instate	27%
y5_p75_earnings	31%	y1_grads_nme	0.31%
y5_grads_earn	31%	y5_grads_nme	0.20%
y10_p25_earnings	20%	y10_grads_nme	0.13%
y10_p50_earnings	20%		
y10_p75_earnings	20%		
y10_grads_earn	20%		
y1_ipeds_count	76%		
y5_ipeds_count	50%		
y10_ipeds_count	33%		

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: 6-digit Office of Post-secondary Education ID (OPEID).
  - Degree level: Associates, Certificate (<1 year, 1-2 years, 2-4 years, post-baccalaureate certificate, post-master's certificate), Bachelors, Masters, Doctoral-Professional Practice and Doctoral-Research/Scholarship.
  - Degree field: For Graduate Earnings, degree field is defined at 4-digit Classification of Instructional Prog (CIP) level for Certificate, Associates, Bachelors, and Doctoral - Professional Practice degree levels, and 2-digit CIP level for all other degree levels. For Employment Flows, degree field is defined at 2-digit CIP level for all degree levels.
  - Graduation cohorts: 3-year cohorts for Bachelor's, 5-year cohorts for all other degree levels.
  - Years after graduation: 1, 5 and 10 years.
  - Geography: National or 9 census divisions.
  - Industry: 2-digit NAICS.

- 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?
  PSEO is created by merging graduation files submitted by educational institutions with administrative data on jobs collected by the Longitudinal Employer-Household Dynamics Program (LEHD). LEHD is updated quarterly, but we cannot find a regular update schedule for PSEO. PSEO was updated annually from 2018 to 2020 and quarterly in 2021. The most recent update was in October 2021.
- 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?
  Seems to be in waves but no published schedule. Variables and records are added and adjusted in each data release. See Data Notices for details.
- What is the time lag between when an event occurs, when it is recorded, and when that data is available to researchers?
  Can be a year or more, depending on the data release schedule.

5. Integrity

- What are the risks to the integrity of this dataset?
  Graduation files are reported by educational institutions but do not include labor market return information. Earnings and employment data mainly comes state unemployment insurance wage records collected via a voluntary federal-state data sharing partnership. Overall, the data is unlikely to be manipulated to distort labor market returns of particular programs or institutions.
- How are data outliers handled? (May be available from published documentation if not metadata.)
  PSEO Graduate Earnings drops graduates who earn less than the annual equivalent of full-time work at the prevailing federal minimum wage. Additionally, graduates with two or more quarters with no earnings in the reference year are also dropped. The data developer believes that these workers are likely to be either marginally attached to the labor force or employed in non-covered employment. In Employment Flows, these workers are categorized as non-employed and assigned a firm industry and geography of “unclassified.”
- 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?
  Data for all institutions and for each state are both downloadable in CSV format. XLS files are also available for state datasets. All past releases are available.
- Are any variables or cases withheld from researchers? If so, does that withholding or censoring

affect researchers' ability to use the data?
Yes. PSEO contains no demographic information and uses differential privacy techniques to ensure that data cannot be used to infer information on individuals. This withholds researchers from using the data to identify disparities by race and gender in NDC attainment.

For earnings and employment variables, values that do not meet Census Bureau publication standards are suppressed and flagged.

- 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, no single variable can serve as a unique identifier. For Graduate Earnings, the combination of aggregate level, institution, degree level, CIP code, and graduation cohort uniquely identifies a record. For Employment Flows, the combination of institution, degree level, CIP code, graduation cohort, geography, and industry uniquely identifies a record.
- 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, both CIP and NAICS codes are available.

8. Suitability for Longitudinal Research?

- Is the metadata consistent over time, at least for key variables?
  Metadata has been adjusted over time as described in Data Notices. This does not significantly affect researchers' ability to use the dataset for longitudinal research as the dataset is already constructed in a longitudinal format and each release includes data from past releases.
- 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?
  Observations overtime for the same unit of analysis (a graduation cohort of a program/institution) are already linked together in the dataset. For each graduation cohort, labor market return information 1/5/10 year(s) after graduation is listed as separate variables of the same record.
- How far back do administrative records from this dataset go? As of the 2021Q3 release, the data covers 2001-2019.

Colorado License Roster (Nursing)

EXECUTIVE SUMMARY

Key Dataset Facts

Dataset Name: **License Roster: Colorado Licensed Practical Nurses**

Publisher: State of Colorado, Department of Regulatory Agencies

Website: https://apps.colorado.gov/dora/licensing/Lookup/GenerateRoster.aspx

Unit of Analysis: Individual

Quality Metrics

Metric	Rating	Summary Explanation
Non-response	Excellent	Very few variables have significant amounts of missing data, and those variables (e.g., zip code) are of limited value for researchers.
Coverage	Excellent	By law, all practical nurses must be licensed. This dataset contains entries for all practical nurses.
Granularity	Good	There isn't a lot of variation in many of the data fields. However, dates and degrees reported are exact, as are addresses.
Consistency	N/A	No basis for comparison to other datasets.
Timeliness	Excellent	Since the data is used for public verification and enforcement, it is updated continuously.
Integrity	Good	It is theoretically possible that individuals may misrepresent data when reporting to the state. However, given the legal implications of falsifying data in the licensure process it is unlikely that this is a significant threat to the data's integrity.
Accessibility	Excellent	Comprehensive microdata can easily be retrieved from the Colorado DORA (Department of Regulatory Agencies) website.
Interoperability	Good	Theoretically, one could use individuals' names and addresses to link to other public records that may be of interest, as is commonly done by market research firms using these rosters.
Suitability for Longitudinal Research	Poor	One would need to have access to archived files (not published online) to see changes in the characteristics of the licensed population over time.
Overall Recommendation	Good	This is a high-quality source of information on licenses in specific occupation(s) that are part of the skilled technical workforce. However, use cases are likely to be limited (see below).

Relevance to Use Cases

Use Case	Rating	Summary Explanation
Analyze the Overall Prevalence of NDCs	Poor	This license roster only covers one specific occupation. In conjunction with other rosters it may allow us to get a clear view of the overall prevalence of licensure.
Identify Which NDCs are Associated with Highest Earnings	Poor	We are unable to say anything about the earnings of nurses using this dataset alone, or in conjunction with other datasets covered in this DQA.
Identify Patterns of Inequality in NDC Attainment	Fair	We may be able to analyze patterns of spatial inequality in the attainment of licensure using this dataset.
Enrichment of NTEWS Microdata	Fair	Name and address data could theoretically be cross-referenced with restricted-use NTEWS variables to validate self-reported data on licenses, or enrich such data (e.g., by importing license start and end dates).

DATA QUALITY ASSESSMENT

Rubric for Assessing Dataset Quality

1. Relevance

- What is the total number of items relevant to non-degree credentials?
  29 variables in total, including the name, address, city/county, state, zip code, and degree of a licensee, and their license number, type, nurse compact designation, first issue date, last renewed date, expiration date, status description, as well as public action case tracking number, program action levied, effective and complete dates of the action.

- What are the measures of NDC attainment like?
  License. Per BLS definition, a license is a credential awarded by a government agency and conveys a legal authority to work in an occupation.

- Are there any indicators related to education attainment that are unique to this dataset?
  License number, first issue/last renewed/expiration dates, status description, and public/program action records are unique to this administrative dataset.

- Are there indicators of other phenomena that could be of sociological significance? No.

- 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 for license lookup and verification.
  a. Measuring the rate of attainment of non-degree credentials within the U.S. skilled technical workforce
     Relevant. Licensee count can be easily obtained from this dataset and used to calculate the attainment rate of a (group of) license within Colorado.
  b. Measuring aggregate returns to non-degree credentials by credential type
     Somewhat relevant. The dataset can be linked to labor market outcome records of the licensees as their names and addresses are given, but such linkage can be challenging as no common unique ID of the licensees (e.g., SSN) is available.

c. Identifying disparities by race and gender in the attainment of non-degree credentials
Somewhat relevant. The dataset contains no race/gender information but can be linked to datasets with such 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 can be linked to labor market outcome records of the licensees.
e. Evaluating the effectiveness of public policies that support the attainment of non-degree credentials?
Relevant. The dataset can be used to study the impact of licensure on the specified occupation.
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 dataset attempts to cover all licensed practical nurses in Colorado.
- What is the number of cases, and how does that number compare to known estimates of the relevant population?
As of June 16, 2022, the dataset has 47,895 records. Since licensure is a government activity and the dataset comes from an administrative database, we believe this number itself represents the best estimate of the population, given there is no significant withholding of license records in the dataset.
- How does the publisher of the data ensure that data is collected for cases that should be in the dataset?
Data is collected through internal system updates, and updates are made by licensees and applicants through online services and by individual board/program actions. Collection instruments include internet update, paper form update and manual data entry.
- Do cases that we believe should exist in the microdata actually exist in the data? We do not see any sign of missing cases in our assessment.
- What percent of cases lack data for key variables of interest? (Direct assessment if not in metadata.)
Formatted name, license type and license number have no missing value. For other key variables, missing rate is minimal except for degrees. Address line 2 and county have high missing rates probably because relevant information is not applicable to these variables. See below for details.

Variable	Missing Count	Missing Pct	Variable	Missing Count	Missing Pct
Formatted name	0	0	License type	2	0
Address line 1	57	0.12	License number	2	0
Address line 2	45,529	95.06	Nurse compact designation	2	0
City	19	0.04	License first issue date	22	0.05
County	35	0.07	License Last Renewed Date	23	0.05
State	25,818	53.91	License Expiration Date	22	0.05
Zip code	29	0.06	License Status Description	2	0

Zip code +4	29	0.06	Degree(s)	14,914	31.14

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?
  - Location: 2-line address, city, state, 9-digit zip code.
  - Degrees: specific degree name.
  - Nurse compact designation: Single state/Multi-state
  - License status: Active/Active – Restricted/Active - With Conditions/Expired/Inactive/Revoked/Summary Suspension/Surrendered/Suspended/Voluntary Surrender/Volunteer.
  - Program action: specific program action type.
- 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?
  Updated in real time.
- 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?
  Continuously. Since the data is not centrally updated, records can 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?
  The dataset is updated in real time.

5. Integrity

- What are the risks to the integrity of this dataset?
  Since licensees and applicants can update certain variables by themselves, they may misreport some information (e.g., degree) in their favor.
- How are data outliers handled? (May be available from published documentation if not metadata.) No relevant information in technical documentation. Outlier is not a significant issue in this dataset as there is no continuous variable.
- 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/CSV/text formats.

- Are any variables or cases withheld from researchers? If so, does that withholding or censoring affect researchers' ability to use the data?
  Three variables (sub-category of license, licensee specialty, and licensee title) have no observations in this dataset. We do not know whether relevant information is withheld or non-applicable.

- 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. License number does not serve as a unique identifier. If a licensee has multiple public disciplinary actions, there will be multiple rows generated for the licensee in order to list each case attached to the license record.

- 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.

- 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?
  No commonly used occupation/industry codes included.

8. Suitability for Longitudinal Research?

- Is the metadata consistent over time, at least for key variables?
  We were unable to determine whether there were significant changes over time in the metadata.

- 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 licensee over time can be linked together through name/location/license number, but historical datasets are not available on Colorado.gov.

- How far back do administrative records from this dataset go?
  We assume that administrative records date back to the establishment of licensure for nurses in Colorado – probably several decades – though it is almost certain that the format of the data has changed over the years.

Maryland Noncredit Workforce Completers System

EXECUTIVE SUMMARY

Key Dataset Facts

Dataset Name: **Maryland Noncredit Workforce Completers System**

Publisher: Maryland Higher Education Commission (MHEC)

Website: https://data.mhec.state.md.us/Documentation_mac2_NWCS.asp

Unit of Analysis: Individual

Purpose of Dataset: To track labor market outcomes associated with the completion of non-credit courses offered by Maryland public higher education institutions. This dataset is included in our data quality assessment as a representative of other states' noncredit education tracking systems.

Quality Metrics

Metric	Rating	Summary Explanation
Non-response	Good	MHEC has access to high quality data on earnings from state data systems. However, there are limitations in Maryland's ability to track students from other states, or those who move out of Maryland shortly after credential completion.
Coverage	Good	Public institutions are required to submit data to MHEC on enrollments. However, private institutions are not required to submit data and, as a small state in a metropolitan area that crosses state boundaries, many Marylanders may be receiving noncredit workforce instruction in states not covered by this dataset.
Granularity	Excellent	Detailed information is available on the noncredit programs completed by students and their future earnings.
Consistency	N/A	We could not identify a suitable comparison dataset. While the data collection format for post-completion earnings is similar to the WIOA PIRL, the population is very different.
Timeliness	Good	Data is updated annually.
Integrity	Excellent	Excellent We did not identify any risks to data integrity.
Accessibility	Fair	While researchers should be able to access microdata, the procedure for data access is unclear.
Interoperability	Good	Researchers could probably link to TrainingProviderResults.gov for some program outcomes.
Suitability for Longitudinal Research	Poor	The dataset is less than one year old (first data was reported in 2021).
Overall Recommendation	Good	This dataset has potential for helping us fill in gaps in our understanding of the value of non-degree credentials.

Relevance to Use Cases

Use Case	Rating	Summary Explanation
Analyze the Overall Prevalence of NDCs	Fair	This dataset is limited to non-credit credentials in one particular state, but does have an impressive level of information on who completes which credentials.
Identify Which NDCs are Associated with Highest Earnings	Good	Data on post-completion earnings is a strength of this dataset.
Identify Patterns of Inequality in NDC Attainment	Good	The dataset contains necessary information on gender and race to identify patterns of inequality in attainment and post-completion value.
Enrichment of NTEWS Microdata	Good	Despite limitations, it (and, presumably, similar systems in other states) could be used to identify the characteristics of non-credit certificate programs reported by individuals responding to the NTEWS.

DATA QUALITY ASSESSMENT

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.

Credential Engine Credential Registry

EXECUTIVE SUMMARY

Key Dataset Facts

Dataset Name: **Credential Registry**

Publisher: Credential Engine

Website: https://credentialfinder.org/

Unit of Analysis: Credential

Purpose of Dataset: To map the competencies associated with all credentials of all types in a way that prospective learners can identify what they should be able to do upon completion, along with creating a repository of other data points about individual credentials.

Quality Metrics

Metric	Rating	Summary Explanation
Non-response	Fair	The majority of registry entries appear to have at least one field missing. Whether non-response is a problem for the researcher would depend on the use case.
Coverage	Poor	According to research commissioned by Credential Engine itself, the registry currently contains about 4% of all U.S. credentials.
Granularity	Excellent	Descriptions of the competencies associated with particular credentials are quite specific.
Consistency	Fair	While it is difficult to compare directly to other datasets, we can note that the coverage of certifications and licenses in the Registry is far more limited than coverage in the DOL Certification Finder and License Finder databases.
Timeliness	Good	Data is released regularly, though institutions can delay providing updates.
Integrity	Excellent	Excellent We did not identify any risks to data integrity.
Accessibility	Fair	While researchers should be able to access microdata, the procedure for data access is not published.
Interoperability	Fair	Linkages could be made to other data sources on particular credentials such as IPEDS and TrainingProviderResults.gov, but the value in such linkages is unclear since the registry's coverage is limited.
Suitability for Longitudinal Research	Good	Some data goes back to 2017.
Overall Recommendation	Fair	While the dataset has promise if it expands in the future, at this time its coverage is too limited to enable nationally-representative analyses.

Relevance to Use Cases

Use Case	Rating	Summary Explanation
Analyze the Overall Prevalence of NDCs	Fair	Since coverage is limited and participants are self-selected, the registry itself does not lend itself to analyzing the overall prevalence of credentials.
Identify Which NDCs are Associated with Highest Earnings	Fair	Through linkages with external data, it may be possible to identify which competencies are associated with labor market returns.
Identify Patterns of Inequality in NDC Attainment	Fair	Since data is at the credential level, one would need to establish linkages with external datasets to identify patterns of inequality.
Enrichment of NTEWS Microdata	Fair	It is possible in theory to import data on the competencies that NTEWS respondents who hold credentials listed in the Registry should have.

DATA QUALITY ASSESSMENT

Rubric for Assessing Dataset Quality

1. Relevance

- What is the total number of items relevant to credentials?
  There are 43 variables in the database's Credential section, including a credential's name, description, type, subject, status, webpage, Credential Transparency Identifier (CTID), language, organization, jurisdiction, audience level, estimated cost, estimated duration, industry type, occupation type, requirements, external quality assurance, online or physical location, financial assistance availability, related credentials, holder number and characteristics, employment and earnings outcome, renewal requirements and frequency, revocation criteria and process, and maintenance process.

- What are the measures of credential attainment like?
  Degree, diploma, license, certificate, certification, courses, and training programs.

- Are there any indicators related to education attainment that are unique to this dataset? Cost, duration, financial assistance, and labor market outcome data of a credential.

- 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 Credential Registry is created by Credential Engine to house up-to-date information about all credentials with a common description language. It is intended to promote credential transparency, enable credential comparability, and support customized search about credentials.
  a. Measuring the rate of attainment of credentials within the U.S. skilled technical workforce
     Relevant. For each credential, it is optional for the credentialing organization to upload the number and characteristics of credentialed individuals and their geographic location. This information, when available, can be used to calculate the rate of attainment of a credential.
  b. Measuring aggregate returns to credentials by credential type
     Relevant. Organizations can choose to upload employment and earnings data of a credential.
  c. Identifying disparities by race and gender in the attainment of credentials
     Relevant. The holders profile includes demographic information.

d. Identifying which credentials are associated with the strongest labor market returns for individuals in the skilled technical workforce
Relevant. Employment and earnings data are available for a credential if the credentialing organization chooses to upload them.
e. Evaluating the effectiveness of public policies that support the attainment of credentials? Relevant. Researchers can use the database to assess the labor market outcome and financial assistance landscape of credentials and help policymakers make relevant supporting policies.

2. Coverage

- What is the frame of reference for the dataset, what population does the dataset attempt to cover? The Credential Registry attempts to cover all credentials in the United States.
- What is the number of cases, and how does that number compare to known estimates of the relevant population?
As of September 2022, the Credential Registry includes 40,890 credentials. Credential Engine identifies nearly one million credentials in the United States, so the database seems to cover about four percent of them.
- How does the publisher of the data ensure that data is collected for cases that should be in the dataset?
Credential Engine depends on credentialing organizations to upload information about their credentials. It also cooperates with 28 states and regions for bulk upload of credential information in their jurisdiction.
- What percent of cases lack data for key variables of interest? (Direct assessment if not in metadata.) For each credential, its name, description, type, status, webpage, CTID, language, and organization are required. Other information is optional. Among all 40,890 credentials, 9,759 have a cost profile, 16,814 have a duration profile, 3,408 have industry information, 14,086 have occupation information, and 1,034 have financial aid information. Coverage for a selection of optional information can be found here.

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?
  - Type: Apprenticeship Certificate/Associate Degree/Bachelor's Degree/Badge/Certificate/Certification/Degree/Digital Badge/Diploma/Doctoral Degree/General Education Development (GED)/Journeyman Certificate/License/Master Certificate/Master's Degree/Micro-Credential/Open Badge/Professional Doctorate/Quality Assurance Credential/Research Doctorate/Secondary School Diploma
  - Status: Active/Deprecated/Probationary/Suspended/Teach Out
  - Audience level: Beginner/Intermediate/Advanced/Lower Division/ Upper Division/ Secondary School or Equivalent/Post-Secondary (Associates Degree/Bachelor's Degree/Master's Degree/Doctoral Degree)/Undergraduate/Graduate/Professional/Remedial
  - Industry type: recommend NAICS code recommended
  - Occupation type: O*NET code recommended
  - Financial assistance: General Financial Assistance (Grant/Loan/Scholarship/Work-

Based)/Military

- 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?
  According to the technical documentation, the update frequency varies with the credential. For each credential published, the entity issuing that credential will be asked for an estimate of how frequently the information about that credential is likely to change. If there have been no updates provided by the owner about the credential in that timeframe, Credential Engine will automatically follow-up with the owner to confirm that the existing information remains accurate. If changes need to be made, the owner, working with Credential Engine, will determine if a new version is warranted or not. If no changes are made, the currency clock on that credential is reset.

- 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?
  Continuously.

- What is the time lag between when an event occurs, when it is recorded, and when that data is available to researchers?
  As the registry is continuously updated, there should be little or no lag between reporting and data availability. However, it is ultimately up to the credential issuers and collaborating state higher ed systems to report data in a timely manner.

5. Integrity

- What are the risks to the integrity of this dataset?
  Information is self-reported by credentialing organizations, which theoretically can have an incentive to misreport the cost, duration, holder, and employment profiles of the credential to attract more students/users. Credential Engine claims to authenticate the information uploaded by organizations.

- 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?
  Organizations with approved Credential Engine accounts can download the Registry data. There is no published application form or process for independent researchers to apply to access the registry microdata beyond looking up individual credentials via Credential Finder. However, researchers and other users are invited to contact Credential Engine to learn about microdata access options.

- Are any variables or cases withheld from researchers? If so, does that withholding or censoring affect

researchers' ability to use the data?
No.

- Are there direct costs or indirect costs (e.g., training, resources) associated with accessing and using the data?
  Yes. It is possible that Credential Engine may charge fees for registry data access depending on the intended/licensed use of the data. There may also be opportunity costs associated with becoming proficient in the use of Credential Transparency Description Language (CTDL) that is used to classify all credentials in the registry.

7. Interoperability

- Is there a unique identifier for individual cases? If so, is it one that can be found in other datasets? Yes, the Credential Transparency Identifier (CTID), assigned by Credential Engine, is unique in the database but does not apply in other 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).
  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?
  Credential Engines recommends that organizations submit NAICS and O*Net codes, but organizations may choose to use other classification systems.

8. Suitability for Longitudinal Research?

- Is the metadata consistent over time, at least for key variables?
  Metadata has been modified over time, but the reporting standards for key variables seem to remain consistent. Metadata release history can be found here.
- 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?
  It seems that each credential has a unique entry, and whether historic versions of the credential can be accessed is unclear.
- How far back do administrative records from this dataset go? 2017.

Integrated Post-secondary Education Data System (IPEDS)

EXECUTIVE SUMMARY

Key Dataset Facts

Dataset Name: **Integrated Post-secondary Education Data System (IPEDS)**

Publisher: U.S. Department of Education

Website: https://nces.ed.gov/ipeds/use-the-data/download-access-database

Unit of Analysis: Credential Issuer & Credential

Quality Metrics

Metric	Rating	Summary Explanation
Non-response	Excellent	Missing data is rare. Institutions must report data to maintain eligibility for federal student aid programs.
Coverage	Good	Covers all for-credit certificates issued by institutions that choose to participate in federal student aid programs. Known weaknesses include the exclusion of non-credit (though proposals exist to expand non-credit data collection) and the limited number of post-secondary institutions that do not participate in federal aid programs.
Granularity	Excellent	Excellent Most data fields are detailed, including highly specific Classification of Instructional Programs (CIP)
Consistency	N/A	No basis for comparison to other datasets.
Timeliness	Good	Updated on an annual basis. Published data on the Department of Education's website should be no more than a year old.
Integrity	Good	It is theoretically possible that institutions may attempt to skew reporting in their own favor. However, given the legal implications of falsifying data, it is unlikely that this is a significant threat to the data's integrity.
Accessibility	Good	Comprehensive microdata can easily be downloaded free of charge. Some researchers have reported difficulty in working with Microsoft Access files.
Interoperability	Excellent	CIP and institution codes in IPEDS can be linked to other datasets where an educational institution can be positively identified. For example, IPEDS could be used to enrich PSEO data with information on an institution's resources, student to faculty ratio, tuition price, and so on.
Suitability for Longitudinal Research	Good	Archived annual IPEDS files can be downloaded. However, one may need to do some data cleaning and analysis to create custom data files containing data points from multiple reporting years.

Metric	Rating	Summary Explanation
Overall Recommendation	Excellent	This is a great data source for learning about the overall landscape of higher education in the United States. While it does not contain data on all credentials, it can be used to enrich other datasets.

Relevance to Use Cases

Use Case	Rating	Summary Explanation
Analyze the Overall Prevalence of NDCs	Good	IPEDS data can be used to estimate the overall number of individuals across the United States enrolled in and graduating from specific degree and certificate programs.
Identify Which NDCs are Associated with Highest Earnings	Fair	While not usable for this purpose in and of itself, IPEDS can potentially be linked to other datasets with earnings data at the institution level (i.e., PSEO).
Identify Patterns of Inequality in NDC Attainment	Good	Data on student demographics can be used to identify whether particular races, ethnicities, or genders complete for-credit credentials at different rates.
Enrichment of NTEWS Microdata	Excellent	IPEDS can be used to import rich data on the educational institutions reported by survey respondents.

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.

7. 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.

Registered Apprenticeship Partners Information Data System (RAPIDS)

EXECUTIVE SUMMARY

Key Dataset Facts

Dataset Name: **RAPIDS (Registered Apprenticeship Partners Information Data System)**

Publisher: U.S. Department of Labor, Office of Apprenticeship (OA)

Website: https://www.dol.gov/agencies/eta/apprenticeship/about/statistics/2021 Unit of Analysis: Worker AND Credential

Quality Metrics

Metric	Rating	Summary Explanation
Non-response	Excellent	Excellent 3,043,210 individuals and 105,584 programs, which is consistent with existing estimates of the number of credentials and programs in existence with the known exclusion of certain states (cover Minnesota, Oregon, Vermont, Washington state, and the District of Columbia). No known missing cases outside of those states.
Coverage	Good	Coverage rates range from 73 to 100% for individual variables except for county, which is filled in for only about 65% of cases.
Granularity	Good	Contains detailed O*Net and NAICS codes and earnings data; however some data is lumped into poorly defined categories (e.g., educational attainment)
Consistency	N/A	We have not been able to find datasets that would lend themselves to an "apples to apples" comparison with RAPIDS.
Timeliness	Good	Data is reported to the website on a regular basis.
Integrity	Excellent	Excellent We did not identify any risks to data integrity.
Accessibility	Good	Anyone can download and work with the microdata immediately. However, the page containing the microdata file is difficult to find within dol.gov.
Interoperability	Good	Data can theoretically be matched to other sources on the name of the apprenticeship program, though no widely accepted, unique identifier for apprenticeship programs exists.
Suitability for Longitudinal Research	Excellent	The most recent edition of RAPIDS contains data on all apprentices in the United States (except MN, OR, VT, WA and DC) going back to 2000.
Overall Recommendation	Good	This is a high-quality source of information on one particular type of non-degree credential (apprenticeship).

Relevance to Use Cases

Use Case	Rating	Summary Explanation
Analyze the Overall Prevalence of NDCs	Excellent	RAPIDS gives us accurate data on the whole universe of apprentices.
Identify Which NDCs are Associated with Highest Earnings	Good	RAPIDS contains enough data on post-apprenticeship wages to tell us about how much apprentices earn while in their programs and upon completion. However, as with PIRL, data on individual earnings is limited to four quarters after program completion.
Identify Patterns of Inequality in NDC Attainment	Excellent	RAPIDS can tell us about the demographic characteristics of apprentices with enough detail to identify disparities in earnings or job quality on the basis of race, even if can't tell us why those disparities occur.
Enrichment of NTEWS Microdata	Fair	NTEWS, while containing data on whether one completed an apprenticeship, does not ask for enough details about respondents' apprenticeship programs to positively match to an apprenticeship program, or an individual apprentice, contained in RAPIDS.

DATA QUALITY ASSESSMENT – REGISTERED APPRENTICESHIP PARTNERS INFORMATION DATABASE SYSTEM (RAPIDS) FY2022 Q1 EXTRACT

Rubric for Assessing Dataset Quality

1. Relevance

- What is the total number of items relevant to non-degree credentials?
  The dataset consists of three tables: All Apprentice, All Program, and Program Occupation.
  - All Apprentice: 66 variables including apprentice ID, age at start, gender, race, ethnicity, education level, indicators for veteran and disabled status, program information, related technical instruction (RTI) provider information, occupation information, registered date, starting date and wage, exit date and wage, expected completion date.
  - All Program: 48 variables including program ID, name, location, type, status, employer type, standards type, registered and updated dates, NAICS and SIC codes, active apprentice count, workforce size.
  - Program Occupation: 37 variables including program ID, name, location, status, occupation title and code, NAICs and SIC codes, RTI length and wage, active apprentice count, number of female/youth/minority/journey workers employed, prevailing journey worker wage,

- What are the measures of NDC attainment like?
  Apprenticeship, defined as an industry-driven career pathway where employers can develop and prepare their future workforce, and individuals can obtain paid work experience, classroom instruction, mentorship, and a portable credential. A Registered Apprenticeship is a proven model of apprenticeship that has been validated by the Department of Labor's Office of Apprenticeship (OA) or a State Apprenticeship Agency (SAA).

- Are there any indicators related to education attainment that are unique to this dataset?
  Yes, individual apprentice records and apprentice counts of a program are unique to this dataset.

- Are there indicators of other phenomena that could be of sociological significance?
  Yes, the dataset contains demographic information of apprentices as well as the employment of female/youth/minority/journey workers in apprenticeships.

- 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.)
  RAPIDS is the primary case management platform for Registered Apprenticeships nationwide. The extract of RAPIDS data, assessed in this document, is made for the public to download and use in research.
  a. Measuring the rate of attainment of non-degree credentials within the U.S. skilled technical workforce
     Relevant. The dataset includes the active apprentice count of a program.
  b. Measuring aggregate returns to non-degree credentials by credential type
     Relevant. The dataset includes starting and exiting wages of apprentices.
  c. Identifying disparities by race and gender in the attainment of non-degree credentials Relevant. The dataset includes apprentice race, ethnicity, and gender as well as the number of females/minorities employed.
  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 apprentice exiting wage but no subsequent earnings data.
  e. Evaluating the effectiveness of public policies that support the attainment of non-degree credentials?
     Relevant. The dataset can be used to evaluate the returns to an apprenticeship, the race and gender disparities in an apprenticeship, and whether/how the apprenticeship should be supported by public policy.

2. Coverage

- What is the frame of reference for the dataset, what population does the dataset attempt to cover?
  RAPIDS attempts to cover all Registered Apprenticeships and their apprentices in the United States.

- What is the number of cases, and how does that number compare to known estimates of the relevant population?
  The FY2022 Q1 release of the dataset contains 3,043,210 individual apprentice records and 105,584 apprenticeship programs. RAPIDS currently captures individual level data for 48 states and the United Services Military Apprenticeship Program (USMAP). It does not include data from Minnesota, Oregon, Vermont, Washington state, and the District of Columbia, but is expected to represent the complete national dataset by the end of FY2022. See here for details.

- How does the publisher of the data ensure that data is collected for cases that should be in the dataset?
  Previously, OA only had access to individual level data from the 41 states that used RAPIDS. The remaining 12 states and territories were SAA states that used their own system and only submitted aggregate data to OA. OA also only received aggregate data for USMAP. In FY2021, OA created a portal that allowed SAA states to transfer non-personally identifiable, individual level data into RAPIDS, including demographic data. RAPIDS now captures individual level data from 48 states and USMAP. OA is working with the five remaining states and territories with the goal to have all national individual

level data by the end of FY2022.

- Do cases that we believe should exist in the microdata actually exist in the data? We did not see any sign of missing cases in our assessment of the data.
- What percent of cases lack data for key variables of interest? (Direct assessment if not in metadata.)

All Apprentice (N=3,043,210)			**All Program** (N=105,584)			**Program Occupation** (N=155,292)		
Variable	**Miss#**	**Miss%**	**Variable**	**Miss#**	**Miss%**	**Variable**	**Miss#**	**Miss %**
Program ID	0	0.0%	Program Name	1,070	1.0%	Program Name	48	0.0%
Program Name	19	0.0%	Program Status	690	0.7%	Program Status	7	0.0%
Program Type	26,741	0.9%	Employer Type	2,560	2.4%	NAICS Code	42,896	27.6%
Program Status	3	0.0%	Standards Type	860	0.8%	SIC Code	32,982	21.2%
RTI Provider Name	556,375	18.3%	NAICS Code	31,451	29.8%	Program Address	8,553	5.5%
Occupation Title	1,683	0.1%	SIC Code	25,903	24.5%	Program City	1,271	0.8%
O*NET Code	67,071	2.2%	Program Address	7,335	6.9%	Program State	47	0.0%
NAICS Code	157,269	5.2%	Program City	1,889	1.8%	Program ZIP	375	0.2%
Term Length Max	4,643	0.2%	Program State	1,069	1.0%	County	55,758	35.9%
Term Length Min	491,531	16.2%	Program ZIP	1,334	1.3%	Region	34,640	22.3%
Apprentice Status	1,109	0.0%	County	43,309	41.0%	Registered Date	323	0.2%
County	431,151	14.2%	Region	1069	1.0%	Updated Date	34,855	22.4%
Registered Date	434,448	14.3%	Registered Date	1,272	1.2%	Occupation Title	5,990	3.9%
Start Date	125,208	4.1%	Updated Date	26,577	25.2%	O*NET Code	14,006	9.0%
Starting Wage	1,585	0.1%	Active Appr. Count	18	0.0%	Occupation Type	7,831	5.0%
Exit Date	557,818	18.3%	Workforce Size	28,292	26.8%	Term Length Max	40,156	25.9%
Exit Wage	329,833	10.8%				Term Length Min	34,041	21.9%
Expected Comp. Date	126,903	4.2%				RTI Length	34,147	22.0%
Age at Start	184,817	6.1%				RTI Length Type	129,877	83.6%
Gender	507	0.0%				RTI Hours	76,426	49.2%
Race	264,260	8.7%				Probationary Length	35,387	22.8%
Ethnicity	507	0.0%				Active Appr. Count	0	0.0%
Education Level	45,933	1.5%				Journeyman Emp. Count	34,394	22.1%

All Apprentice (N=3,043,210)			**All Program** (N=105,584)			**Program Occupation** (N=155,292)		
Variable	**Miss#**	**Miss%**	**Variable**	**Miss#**	**Miss%**	**Variable**	**Miss#**	**Miss %**
Veteran Status	507	0.0%				Journeyman Wage	41,593	26.8%
Disabled Status	507	0.0%				Female Emp. Count	39,066	25.2%
						Youth Emp. Count	39,065	25.2%
						Minority Emp. Count	39,067	25.2%

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?
  - Apprentice status: Registered/Completed/Transferred/Suspended/Reinstated/Cancelled
  - Gender: Female/Male/Not provided
  - Race: American Indian or Alaska Native/Asian/Black or African American/Native Hawaiian or other Pacific Islander/White/Multiple-race selected/Do not wish to answer
  - Ethnicity: Hispanic/Non-Hispanic/Not provided
  - Education Level: 0/1/2/3/4/5/6/7/8/9

    0 - Unknown
    1 - Not High School graduate
    3 - High School graduate (including equivalency)
    4 - GED (legacy system)
    5 - Unknown (legacy system)
    6 - Some College or Associate's degree
    7 - Bachelor's Degree
    8 - Master's Degree
    9 - Doctorate or Prof. degree
  - Veteran Status: Yes/No/Not provided
  - Disabled Status: Yes/No/Not provided
  - Program type: 0/1/2/3/4 1- Individual Non-Joint (INJ) 2- Individual Joint (IJ) 3-Group Joint (GJ) 4-Group Non-Joint (GNJ)
  - Program status: Registered/Registration pending approval/Incomplete registration/Suspended/Re-instated/Pending cancellation/Cancelled/Deleted
  - Program location: address, city/county, state, and 5-digit ZIP
  - Occupation title: specific occupation
  - Occupation type: Competency-based/Time-based/Hybrid
  - Industry and occupation codes: 6-digit NAICS, 4-digit SIC, 8-digit O*NET
  - Employer type: Single employer/Multi-employer/NA
  - Standards type: National program standards/Local Apprenticeship standards/NA
  - RTI length type: A/T (Annual/Full term of apprenticeship)
- 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?
  A new file is posted to the OA website every quarter one month after the last day of the quarter, with some variation depending on the extent of data cleaning required.

- 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?
  Although the PUF is updated quarterly, data comes in continuously and is updated as submitted by sponsors. In some states, the state acts as an intermediary. SAA states (about half of the U.S.) are required to submit quarterly data to OA. Of these, 12 do not use RAPIDS as their case management system. DOL works with them to get their data uploaded into RAPIDS. WA and DC do not yet provide data; VT and MN are in the process of being incorporated.

- What is the time lag between when an event occurs, when it is recorded, and when that data is available to researchers?
  Any event that occurs should be in the PUF within four months; however, OA sees it as soon as submitted.

5. Integrity

- What are the risks to the integrity of this dataset?
  Employers or apprentices may misreport some information, such as wage and program length, in their favor.

- How are data outliers handled? (May be available from published documentation if not metadata.)
  There are zero or negative values in several continuous variables, such as term length, wage, active apprentice count, and workforce size. The technical documentation does not explain why they exist. OA is working to clean up data entry errors, some of which occurred many years ago. Older data is more likely to have null values.

- 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, split into five Excel files, can be downloaded here from the Department of Labor's website.

- Are any variables or cases withheld from researchers? If so, does that withholding or censoring affect researchers' ability to use the data?
  Yes. All personally identifiable information, such as apprentice name and SSN, has been removed. This

may create challenges for linking the dataset to other data sources with individual level 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.

7. Interoperability

- Is there a unique identifier for individual cases? If so, is it one that can be found in other datasets?
  Yes, there are apprentice and program IDs.

- 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, they are unique to RAPIDS.

- 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 SIC codes are available. Within RAPIDS, records can be linked easily with apprentice/program ID. Linking apprentice data to other data sources can be challenging since all personally identifiable information is removed. With O*NET, NAICS, and SIC codes available, linkage to industry/occupation level data is easy.

  DOL/ETA noted that there have been efforts internally to develop a PIRL-RAPIDS linked data file, however data quality issues with PIRL and insufficient resources have been barriers to creating such a dataset.

8. Suitability for Longitudinal Research?

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

  Variables in the data extract are not fully consistent with those in the data dictionary, so there seems to be changes in the metadata.

  In addition, OA indicates that it introduced Total Apprentices Served to show a more comprehensive picture of Registered Apprenticeship system activity. It previously used the Active Apprentice Count to show the overall capacity and scale of the apprenticeship system. The FY2022 Q1 release of the dataset has not included Total Apprentices Served yet but includes Active Apprentice Count only.

  There have been a few changes in response to legislative requirements such as the Veterans Act and other requests in the PRA process. At one point the option to not self-identify was added, possibly July 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 consists of three cross-sectional tables. Observations for the same unit of analysis can be linked together through unique IDs.

- How far back do administrative records from this dataset go?
  The extract includes data on all apprentices, programs, and occupations in the system dating back to 2000.