A Framework for Sampling and Validating Performance Data Submitted Under the Workforce Innovation and Opportunity Act*

Nigel Soria†

February 2020

*This draft is for discussion purposes. Click here for the most recent version of this document.

†Senior Research Economist, Kansas Department of Commerce, 1000 S.W. Jackson Street, Suite 100, Topeka, KS 66612; telephone: (785) 296-1705; e-mail address: nigel.soria@ks.gov.

Summary

Data validation involves using a series of internal controls or quality assurance techniques to verify the accuracy, validity, and reliability of data. This document aims to establish a joint data validation framework based on guidance from the U.S. Departments of Labor and Education. Specifically, this document describes procedures for sampling and validating performance data submitted under the Workforce Innovation and Opportunity Act. The sampling procedure involves selecting records using a two-stage stratified sample design, where workforce investment boards form the first-stage strata and employment service offices constitute the second-stage primary sampling units. Primary sampling units are selected using a probability proportional to size systematic sampling procedure, and individual records are randomly sampled from the selected primary sampling units.

Contents

1 Introduction

Data validation is a series of internal controls or quality assurance techniques established to verify the accuracy, validity, and reliability of data. The purpose of a database audit is to prove that data have been correctly entered from the source document into the computer database; however, the audit does not address the issue that what's contained in the source document could be wrong, which is of course important and must be otherwise taken care of. Establishing a joint data validation framework will ensure that all program data are consistent and accurately reflect the performance of each core program.

1.1 Why Sample (Randomly)?

A sample of the data is audited for the practical reason that a 100 percent audit would take too much time. By selecting a representative subset we can estimate the proportion of data values in error for the database as a whole. There is a straightforward cost/benefit relationship between the amount of time spent performing an audit and the variability of the resulting statistics. Precise estimates, though, do nothing to improve the quality of the data in question. Beyond a certain point, it is more cost-effective to spend that time reviewing the data in the first place rather than counting errors afterwards. Why Sample Randomly? The random selection of a sample ensures coverage of the whole database. Each data value has an equal chance of being chosen for the audit. It is this even-handed random selection process that makes the sample statistics representative of the database as a whole.

2 Background

Section 116 of WIOA establishes performance accountability indicators and performance reporting requirements to assess the effectiveness of states and local areas in achieving positive outcomes for individuals served by the workforce development system's six core programs. These six core programs are the Adult program, Dislocated Worker program, and Youth program, authorized under WIOA title I and administered by DOL; the Adult Education and Family Literacy Act (AEFLA) program, authorized under WIOA title II and administered by ED; the Employment Service program authorized under the WagnerPeyser Act, as amended by WIOA title III and administered by DOL; and the Vocational Rehabilitation (VR) program authorized under title I of the Rehabilitation Act of 1973, as amended by WIOA title IV

and administered by ED. WIOA provides a historic opportunity to align performance-related definitions, streamline performance indicators, integrate reporting, and ensure comparable data collection and reporting across all six core programs, while also requiring the collection and reporting of program-specific data.

Through the guidance of Training and Employment Guidance Letter (TEGL) No. 07-18, the U.S. Departments of Labor and Education elaborate on the performance accountability guidelines required to be developed under WIOA section 116. This guidance provides states with a general framework for data validation. Specifically, the U.S. Departments of Labor and Education have developed this guidance pursuant to WIOA section 116(d)(5), which requires the Departments to establish data validation guidelines to ensure the information contained in program reports is valid and reliable. States must develop data validation procedures consistent with these guidelines.

2.1 Flexible Data Validation Framework

While states must utilize a data validation strategy, the specific design, implementation, and periodic evaluation of that strategy is left to the discretion of the state so long as those strategies or procedures are consistent with federal guidelines. Data validation helps ensure the accuracy of the annual statewide performance reports, safeguards data integrity, and promotes the timely resolution of data anomalies and inaccuracies. As such, it is recommended that states incorporate their data validation procedures into their internal controls procedures, which are required by 2 Code of Federal Regulations (CFR) §200.303. State VR agencies should also consider related guidance issued in Rehabilitative Services Administration (RSA) Policy Directive 16-04. Each state must develop data validation procedures that include:

- written procedures for data validation that contain a description of the process for identifying and correcting errors or missing data, which may include electronic data checks;
- regular data validation training for appropriate program staff (e.g., at least annually);
- monitoring protocols, consistent with 2 CFR §200.328, to ensure that program staff are following the written data validation procedures and take appropriate corrective action if those procedures are not being followed;
- a regular review of program data (e.g., quarterly) for errors, missing data, out-of-range values, and anomalies;

- documentation that missing and erroneous data identified during the review process have been corrected; and
- regular assessment of the effectiveness of the data validation process (e.g., at least annually) and revisions to that process as needed.

2.2 Common Data Elements

Procedures developed by the States must include regular data element validation through core program monitoring on 24 common data elements. The U.S. Departments of Labor and Education selected these elements based on their importance to reporting accurate performance outcomes and to ensure data consistency across core programs. States are encouraged to implement a sampling methodology of their participant files and conduct file reviews of data elements against source documentation. In Attachment I of TEGL 07-18, the U.S. Departments of Labor and Education identify acceptable source documentation necessary to validate these selected data elements. States may: (1) maintain supporting documentation for program-specific data elements not included in this joint guidance; (2) conduct additional source document validation on more data elements; and (3) require additional source documentation in their procedures.[1]

3 Sample Design and Implementation

The sample design used to select records from Participant Individual Record Layout files is complex, so it is very important analysts understand the design in order to apply appropriate techniques and procedures, including error rate estimation. This section describes the design of the PIRL sample, including the target population, design, and processes used to select records from individual PIRL files.

3.1 Target Population

The target population consists of exiters from each of the following funding streams for a particular program year: Adult, Dislocated Worker, Youth, Wagner-Peyser (including JVSG), and TAA. An "exiter" is defined as an individual who has gone at least 90 days without service and does not have anything scheduled at the time of sampling. The sampling procedure excludes

[1]Please refer to Training and Employment Guidance Letter No. 07-18 for additional information on the validation guidelines including a list of the common data elements and the permissible source documentation for each element.

reportable individuals, where a "reportable" individual is someone who has not met the criteria of a participant.

3.2 Sample Design

Each sample is selected using a two-stage stratified sample design. At the first stage, a state is divide into (up to) six sampling strata. Five of the strata correspond to the five WIBs with the largest number of records, and the sixth stratum consists of the remaining WIBs. If a state has five or fewer WIBs, the number of sampling strata equals the number of WIBs. Within each of these strata, primary sampling units (PSUs) are formed and selected. The PSUs for the sample are defined as employment service offices. The sample PSUs are randomly selected using a probability-proportionate-to-size (PPS) procedure that gives a higher chance of selection to PSUs having a larger number of files. To counterbalance this propensity to select offices having the largest caseloads, the sampling scheme selects the same number of records within each PSU regardless of PSU size. In this manner, a record that is created during the sampling period will be included in the sample with approximately equal probabilities within sampling strata regardless of the relative size of the PSU.

3.3 Selection of Offices

Six mutually exclusive and exhaustive categories of records are of interest for the purposes of data validation. They form the within PSU sampling strata, which will be referred to in this document as sampling domains to avoid confusion with the sampling strata formed during the first-stage selection process. The six within PSU sampling domains are described in Section 3.4. To select the sample of PSUs, each PSU in the population is assigned a size measure that is a function of the desired sampling rate for each of the six domains within each stratum and the estimated target population size in each PSU.

The composite size measure for each PSU is computed as follows. There are up to six strata for PSU selection, and within each selected PSU, there are six second-stage strata (or domains). Then, the composite size measure for the ith PSU in the hth stratum is calculated as follows:

$$S_{hi} = \sum_{d=1}^{6} f_{hd} N_{hid},$$

where f_{hd} is the sampling rate for the dth domain within the ith PSU, in the hth first-stage

Table 1: The Six Within-PSU Sampling Domains

Domain	Employment Data	Wage Data	Positive-Valued Elements
1	+	+	+
2	+	+	-
3	+	-	+
4	+	-	-
5	-	-	+
6	-	-	-

stratum. The probability of selection of the ith PSU in the hth first-stage stratum is

$$\pi_{hi} = n_h \times \frac{S_{hi}}{S_{h+}},$$

where $S_{h+} = \sum_i S_{hi}$ is the total size measure of all PSUs in the hth first-stage stratum and n_h is the desired sample size of PSUs from stratum h. An independent sample is then drawn from each first-stage stratum with probability proportional to size (PPS) using systematic sampling.

3.4 Sampling Domains

Essentially, the domain structure consists of the cross-classification of three characteristics. At the first level, records are divided into "contains employment data" (Domains 1-4) and "does not contain employment data" (Domains 5 and 6). The group "contains employment data" is further subdivided into two subdomains corresponding to records that have wage data (Domains 1 and 2) and records that do not have wage data (Domains 3 and 4). Table 1 graphically outlines the domain structure, where "+" indicates the domain has the characteristic in question and "-" indicates the domain does not have the characteristic.

4 Weighting and Error Rate Estimation

This section describes the process used to construct the analysis weight for the PIRL sample. The weight is constructed in stages corresponding to the stages of the sample design described in Section 3.

4.1 Calculation of Analysis Weights

The analysis weight for a record in a PIRL file is the inverse of the probability of selection of the record. The purpose of the weight is to adjust the estimates for differential probabilities

that resulted in the sampling process. The probability of selection for a record within a PIRL file is the product of two probabilities: the first stage selection probability and the second stage selection probability. The first stage probability is the probability of selecting the PSU (office) associated with the record, and the second stage probability is the probability of selecting the record given the record's office is sampled. The inverse of the first stage probability is called the first stage base weight, and the inverse of the second probability is called the second stage base weight.

Before we proceed, we need to define some notation. Specifically, let

h	denote the sampling stratum for the primary sample selection where...;
n_h	denote the number of PSUs sampled in stratum h;
i	denote the PSU sampled within stratum h (i.e., $i = 1, \ldots, n_h$);
d	denote the sampling domain within each PSU, where $d = 1, \ldots, 6$;
n_{hid}	denote the number of records sampled in stratum h, PSU i, and domain d;
j	denote the record sampled within PSU and domain, where $j = 1, \ldots, n_{hid}$;
N_{hid}	denote the number of eligible records in stratum h, PSU i, and domain d; and
π_{hi}	denote the probability of selection for the ith PSU in stratum h.

Finally, let π_{hidj} denote the probability of selection for the jth record in domain d, PSU i, and stratum h. Then,

$$\pi_{hidj} = \pi_{hi} \times \frac{n_{hid}}{N_{hid}},$$

which is the product of the first-stage and second-stage selection probabilities, denotes the overall selection probability of the record. The inverse of this selection probability represents the analysis weight for all records j in stratum h, PSU i, and domain d:

$$W_{A,hidj} = W_{1,hidj} \times W_{2,hidj},$$

where $W_{1,hidj} = 1/\pi_{hi}$ is the first-stage weight and $W_{2,hidj} = N_{hid}/n_{hid}$ is the second-stage weight

4.2 Error Rate Estimation

Forthcoming.