SUPPORTING STATEMENT FOR REQUEST FOR OMB APPROVAL UNDER THE PAPERWORK REDUCTION ACT

PART B – COLLECTION OF INFORMATION EMPLOYING STATISTICAL METHODS

---

1. Description of Universe and Selection Methods Used

As described in Part A of the Supporting Statement, the data validation methodology consists of two parts:

1) **Report validation.** ETA assures the validity of the Senior Community Service Employment Program (SCSEP) aggregate reports by using the SCSEP Performance and Reporting System to automatically generate the grantee-level aggregate reports based on the grantee's individual record files entered into the system and the performance reporting specifications for the quarterly and final year-end report. Edits built into the system assure the validity of SCSEP's performance reports.

2) **Data element validation** assesses the accuracy of participant data records. For SCSEP, the universe for data element validation comprises all SCSEP records submitted to ETA during the prior program year. This information is drawn from national case management software and is not uploaded at the grantee or sub-grantee level. The universe from which the data is obtained for these programs is described in the People Served by Program report available at doleta.gov/Performance/results/pdf/DOL_Workforce_Rprt_Dec_2016.pdf. Data element validation is performed by reviewing samples of participant records against source documentation to ensure compliance with federal definitions and to verify the accuracy of the information contained in the system.

The data validation process results in an estimate of the error rates for each data element and each reported count. Error rates are estimated separately for each grantee for SCSEP.

The methodology for data element validation employs sampling to improve the efficiency of the validation process. To minimize grantees' burden in performing validation consistent with producing a reliable estimate of the error rates, the data element validation process is designed to compute a reliable error rate using the smallest possible sample size. To accomplish these objectives, two sampling techniques are used:

- Variable sampling rates among grantees are used to reduce the burden on small grantees as much as possible.
- Oversampling of high-risk and high-importance cases is used to provide a more accurate estimate of the error rate.

These sampling methods consider the numbers of records so that the overall burden is reduced as much as possible, while still achieving a reliable estimate of error. See the Methodology Details documentation below for more specific details.

To reduce the burden on grantees, the SCSEP Performance and Reporting System includes a validation system that calculates the validation values, imports the reported counts, draws the data element validation samples, produces online and paper validation worksheets, calculates error rates, and produces the validation reports.

Data validation relies on existing records from grantee data in the system and case files. Response rates are not an issue with the data validation process because the data are sampled from the entire participant file and so do not require survey responses.

2. Procedures for the Collection of Information

A. Statistical Methodology for Stratification and Sample Selection

As noted above, report validation does not require grantees to obtain information via surveys because the entire participant file is utilized during this validation procedure.

For data element validation, multistage samples of participant records are drawn. Two independent samples are selected for each grantee: eligibility and performance. Stratification is not employed within the samples in the selection of grantees or records. Records are randomly sampled directly for the eligibility sample, with no weighting. For the performance sample, each record has a probability of selection proportional to its weight.

To increase the efficiency of the process, records in the performance sample receive a risk weight of 1, 2, or 3 based upon two factors: whether the record is a success for calculating performance (i.e., whether the SCSEP participant was employed in the first quarter after exit), and the risk that the data used to calculate performance are in error.

SCSEP Methodological Details

The SCSEP sampling algorithms draw a weighted performance sample from those participants who exited during the retention time period. Each record in the retention cohort is assigned a weight of 1, 2, or 3. Records are sampled with a probability proportional to their size. Thus, instead of selecting records directly as is done with the eligibility sample, records are selected based upon their weight. The details below explain how to draw the SCSEP validation sample. It provides methodological details on the sample frame (i.e. the universe of records from which the sample will be drawn), the sample size, the weighting scheme, and the selection process.
IMPORTANT NOTE: Several of the variables developed during the sampling process are necessary to calculate the error rates for data element validation. The following fields are particularly important for the calculation: each record's weight (whether or not the

record was sampled), each record R_Sel_i value, n (the calculated samples size), and n_Non_Cert. In addition, other variables are needed to test the sampling algorithm.

Calculate the number of enrollments where the participant exited in the retention cohort

N = Count of enrollments where the 4th quarter after the exit quarter is within the report period **and** (PROJECT_TRANSFER_IND <> "Y" **or** PROJECT_TRANSFER_IND is null) **and** PRIOR_GRANTEE_CODE is null **and** TRANSFERRED_DATE is null **and** NON_EXIT_REASON <> ii_Transferred_grantee" **and** TRANSFERRED_GRANTEE_CODE is null **and** SUB_GRANTEE_CODE <> "RG999" **and** have zero reject errors on the DQR associated with it, its participant record, or any of its CSA or UE records.

Note: A retention exiter is any enrollment where the 4th quarter after the exit quarter is within the report period.

Calculate the Sample Size

$$n = \left(\frac{HL}{deff \times t^2 \times P(1-P)} + \frac{1}{N} \right)^{-1}$$

Where n = sample size, HL is the half-length of the confidence interval, N is the number or records, t = 1.96, $deff$ = 2 and P = 0.95. If n is not an integer the number is rounded up to the nearest integer > n.

If N_{fs} >= 500, then CI = 0.035, Else CI = 0.04
If n > 250, set the sample size = 250.

Assigning Risk Weights to Each Record

For each retention exiter:

Set Risk weight = 3 if this exiter has a UE record with FIRST_QTR_WAGES_TEXT = "ii_Yes_in-state" **OR** "iii_Yes_out-of-state" **OR** "iv_Yes_both_in_and_out" **OR** "v_Yes_other_admin" **OR** "vi_Yes_supplemental"
AND
This exiter has a UE record with SECOND_QTR_WAGES_TEXT = "ii_Yes_in-state" **OR** "iii_Yes_out-of-state" **OR** "iv_Yes_both_in_and_out" **OR** "v_Yes_other_admin" **OR** "vi_Yes_supplemental" **OR** (SECOND_QTR_WAGES_TEXT = null and SECOND_QTR_WAGES_AMT > 0)
AND

This exiter has a UE record with THIRD_QTR_WAGES_TEXT = “ii_Yes_in-state” **OR** “iii_Yes_out-of-state” **OR** “iv_Yes_both_in_and_out” **OR** “v_Yes_other_admin” **OR** “vi_Yes_supplemental” **OR** (THIRD_QTR_WAGES_TEXT = null and THIRD_QTR_WAGES_AMT > 0)

Else set Risk Weight = 2 if this exiter has a UE record with FIRST_QTR_WAGES_TEXT = “ii_Yes_in-state” **OR** “iii_Yes_out-of-state” **OR** “iv_Yes_both_in_and_out” **OR** “v_Yes_other_admin” **OR** “vi_Yes_supplemental”
Else set Risk Weight = 1

Set key variables for record selection algorithm

Total_Weight = Sum of the Risk weight for all the retention exiters.
Sampling_Interval = Total_Weight/n.
R_Sel_i is used to identify records that have been selected for the sample, default and initial value = null.

Select first round of records with certainty

Select all records where weight of record >= Sampling_Interval. If this would result in selecting more than n records, randomly select n records where weight of record >= Sampling_Interval. Set R_Sel_i = “Certainty” for all records selected in this step. Select all records where weight of record >= Sampling_Interval. If this would result in selecting more than n records, randomly select n records where weight of record >= Sampling_Interval. Set R_Sel_i = “Certainty” for all records selected in this step.

Calculate variables related to the second round of sampling

Total_Weight_Non_Cert = Sum of the Risk Weight for all of the retention exiters whose weight is less than Sampling_Interval.
n_Non_Cert = n – (number of records with weight >= Sampling_Interval).
Sampling_Interval_Non_Cert = Total_Weight_Non_Cert / n_Non_Cert.

Select second round of records with certainty

Select all records where (risk weight >= Sampling_Interval_Non_Cert and R_Seli is null). If this would result in selecting more than n_Non_Cert records, randomly select records with weight 2 until n_Non_Cert records are selected. For all records selected in this step, set R_Seli = “Certainty”.

If certainty records are selected, update and recalculate variables.

If (number of records with weight >= Sampling_Interval_Non_Cert **and** weight < Sampling_Interval) = 0, skip step 13. If not, set: Total_Weight_Non_Cert = Sum of the

Risk Weight for all of the retention exiters whose weight is less than Sampling_Interval_Non_Cert, n_Non_Cert = n – number of records already selected Sampling_Interval_Non_Cert = Total_Weight_Non_Cert / n_Non_Cert.

Set random variable (needed to sample records)

Random number (R) = Random number between 0 and Sampling_Interval_Non_Cert.

Set variables used to sample records with non-certainty

Counter = 0, which is used to track the sum of the weight of records.
I_Select = R, to track the weight interval used to select records.
i = 1, a subscript to identify the record(s) under consideration.
Num_Sampled = 0, to count the number of records that have been sampled.
w_i, to identify the weight for the *i*th record.

<u>If all records are to be sampled, select all records</u>

If n_Non_Cert >= number of records not sampled, select all records, and set R_Sel_i = "Certainty" for all records.

<u>If the number of records sampled is less than the number of records to be sampled with non-certainty</u>

Do while Num_Sampled < n_Non_Cert

a. If the current record has already been selected for sampling, go to the next record:

If (R_Sel_i is not null) then
    i = i + 1
EndIf

b. If the current record contains the I_Select weighted unit, select the record, up the number of records sampled, up the I_Select weighted unit, up the weight counter and go to the next record.

Else If I_Select >=Counter and I_Select < (Counter + w_i) then
    R_Sel_i = "Random"
    Num_Sampled = Num_Sampled + 1
    I_Select = I_Select + Sampling_Interval_Non_Cert
    Counter = Counter + w_i
    i = i + 1
EndIf
c. If the I_Select weighted unit is less than the counter, then update the I_Select.

Else If I_Select < Counter then
    I_Select = I_Select + Sampling_Interval_Non_Cert
EndIf

d. If the I_select unit is greater than or equal to the counter plus the next record, up the counter and the record.

Else If I_Select >= Counter+ w_i then
    Counter = Counter + w_i
    i = i + 1
EndIf

B. <u>Estimation Procedure</u>

Estimation encompasses computing sample weights and error rates. Validators compare the data from the samples to source documentation. Once all the data have been evaluated, error rates are calculated for each data element. These error rates are estimated using data weighted to account for differences in probability of selection. The validation software computes the sampling errors for each grantee, taking into account the multistage design and the use of unequal weights.

<u>SCSEP Error Rate Calculation</u>

The first step to calculating the weight is to determine the probability of selection for each record. This is used to calculate the record's error weight, which determines how it impacts the error rate calculations. To calculate the error rates, the weights of the records that are in error are divided by the weights of all records validated, or all records sampled depending on the type of error rate calculation. Error rates are calculated for each data element.

For each record, set $Error\_w_{(j)} = 1/p\_selection_i$, $p\_selection_i = 1$ if (R_sel_i = Certainty), else

$$\frac{\text{Risk weight of record} * \text{n\_non\_cert}}{\sum \text{risk weight for all records where R\_sel = Random or R\_sel is null}}$$

Two error rates must be calculated for each data element validated. The numerator is the same for both – the sum of the error weights for those records for which the appropriate data element failed. The denominators differ. For the reported data error rate, the denominator equals the sum of the error weights for all records sampled for the funding stream that should be validated. On the other hand, the overall error rate denominator equals the sum of the error weights for all records sampled for the funding stream. Because users can constantly change their validation results, the results need to be calculated when opened.

REPORTED DATA ERROR RATE = Σ (Error_w(j) * P_FDE) for each record sampled/ Σ (Error_w(j) * VAL(j)) for each record sampled, where P_FDE = 1 if the element failed the validation and P_FDE = 0 if the element passed the validation or if the grantee was not required to validate the element for this record.

VAL(j) = 1 if the grantee was required to validate the element for the record. VAL(j) = 0 if the grantee was not required to validate the element for the record.

OVERALL ERROR RATE = Σ (Error_w(j) * P_FDE) for each record sampled for the / Σ Error_w(j) for each record sampled, where P_FDE = 1 if the element failed the validation and P_FDE = 0 if the element passed the validation or if the grantee was not required to validate the element for this record.

Note that each data element gets its own error rate calculation.

C. Degree of Accuracy Needed for Purpose Described in the Justification

Error rates for each data element have confidence intervals varying with the size of the sample, from 3.5 percent to 4 percent.

D. Unusual Problems Requiring Specialized Sampling Procedures

The discussion above indicates that the methodology uses specialized sampling procedures. The rationale for using these procedures rather than pure stochastic methods is to minimize the burden that data element validation imposes upon the grantees.

3. Response Rates

As mentioned in Part 1, response rate issues do not arise in the data validation program. Data validation relies on existing records from the SCSEP Performance and Reporting System and case files. Through the use of valid sampling techniques, the validation process results in estimates of data accuracy that can be generalized to the universe of data reported to ETA on program performance and activities.

4. Tests of Procedures or Methods

SCSEP has been conducting validation for ten years. The grantees received training prior to beginning validation and receive ongoing training and technical assistance from ETA's data validation contractor and national office staff throughout the validation process. Results of these data validation activities indicate that the methodology has functioned as intended and has enabled states to identify and address reporting errors.

5. Individuals Consulted on Statistical Aspects of the Design

William S. Borden Senior Fellow Mathematica Policy Research, Inc. (609) 275-2321	Donsig Jang Vice President and Director NORC at the University of Chicago (301) 634-9415