**2016 Federal Employee Viewpoint Survey**

**Public Release Data File**

**Compressed File Contents:**

* fevs\_2016\_prdf.csv – Comma-delimited (CSV) Federal Employee Viewpoint Survey (FEVS) data file.
* 2016 PRDF Codebook.xlsx – Excel file containing information about the data.
* 2016 PRDF Basic Labeling SPSS Syntax – SPSS syntax file to load the csv data file and apply labels
* 2016 PRDF Basic Labeling SAS Program – SPSS program file to load the csv data file and apply labels
* Read Me.docx – Information about the 2016 data file

**What’s New in 2016?**

* In previous years, SPSS and SAS versions of the data were prepared and made available. This year only a CSV file is available. CSV-format files are easy to open across a multitude of data processing software (SPSS, SAS, Stata, R, etc.). SPSS and SAS files can be produced with the enclosed syntax files.
* A new data disclosure avoidance method has been applied to this year’s file. Details on page 2.

**Data File Summary**

The public release data file (PRDF) is broken down into four categories:

1. Survey responses (Q1 – Q71): These data are unaltered from the full dataset used to generate reports for agencies and the public. Responses to the Telework and other Work-Life survey items are not included this year to protect the confidentiality of respondents.
2. Demographics (all variables starting with “D”): This year’s public use file only includes five demographics (sex, minority status, supervisory status, age, and planning to leave) and all but sex were collapsed from the full dataset to help protect confidentiality of respondents.
   1. Minority Status (DMINORITY): A combination of the race/national origin and the ethnicity demographics. Those who identify as both White and Non-Hispanic are coded as “Non-minority” and all other combination of responses are coded as “Minority.”
   2. Supervisory Status (DSUPER): Non-Supervisor/Team Leader and Supervisor/Manager/Executive.
   3. Age Group (DAGEGRP): Collapsed to those who are under the age of 40 and those who are 40 and older.
   4. Planning to Leave (DLEAVING): Recoded “Yes, to Retire” to merge them with the “Yes, Other” category.
3. Work Unit (agency and PLEVEL1 variables): These variables indicate where someone works. To protect respondent confidentiality, only work units one level below agency and with a minimum respondent count of 300 are identified.
4. Other: Includes the statistical weighting variable (POSTWT) and a randomly generated ID number (RANDOMID).

**Data Masking Methodology for Disclosure Avoidance**

The 2016 PRDF uses a new method to identify at-risk individuals and an optimized masking process to greatly reduce the risk of re-identification and disclosure of confidential survey responses while maximizing the amount of demographic data that can be kept intact.

When it comes to re-identifying individuals, there are two key pieces of information: where they work and their demographic profile.

The first task is to limit identifiable work units. Only agency and one level below the agency were included, and only for components with a minimum respondent count of 300 were considered. Testing showed this number was an acceptable medium between being able to report more work units while keeping most of the demographic data intact.

The second task in the disclosure avoidance process is to limit the demographic information by reducing the number of demographic variables included in the file and collapsing response choices of those that remain. The fewer distinctions in the demographic information, the less masking that must be performed to hide small groups that are particularly at risk for disclosure. This is accomplished by collapsing response choices together in a logical way, such as combining the original age categories into a simplified under 40/40 and older groups.

The third task is to identify people who are at-risk of being identified. Individuals are stratified into groups by combining their demographic responses together into a string of characters\*. Example:

|  |  |  |
| --- | --- | --- |
| **Example Demographic Profile** | | |
| SEX | (**B**) Female | Combined:  **B B X B** |
| AGE | (**B**) 40 or Over |
| MINORITY | (**X**) Missing |
| SUPERVISOR | (**B**) Supervisor/Manager/Executive |

Everyone in the same work unit who has a profile of **BBXB** would be part of what is called a “cell” that identifies them as having a unique combination identifying characteristics. The FEVS uses a Rule of Ten to protect respondent confidentiality – at least 10 responses are required to produce a report for any work unit. This same rule is applied to the public release data file – any cell with fewer than 10 respondents is considered at risk of disclosure.

The fourth task involves masking the demographic data in an attempt to roll the at-risk cells into larger cells that aren’t at-risk. This is accomplished by systematically setting demographic values (such A or B) to missing (using the dummy value “X”). A demonstration of this masking/substitution procedure is provided on page 3.

*\*For missing demographic data, a dummy value “X” is used.*

**Masking Procedure Demonstration**

In the first pass three at-risk cells are identified (marked in red). Four possible substitutions are presented by replacing one of the demographic values in sequence. For the first at-risk cell (AAAA), changing the fourth “A” value to the “X” value matches the sequence of the AAAX cell which is not at-risk. Everyone in cell AAAA will be reassigned to cell AAAX at the end of this pass through the data. For the at-risk cells ABAB and BABA, a single substitution will not move either into a not-at-risk cell, so not treatment is applied.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Pass 1 (Single Substitution) | | | | | | |
| CELL | COUNT | x--- | -x-- | --x- | ---x | SOLUTION |
| AAAA | 3 | XAAA | AXAA | AAXA | AAAX | **AAAX** |
| AAAX | 13 | -- | -- | -- | -- | -- |
| ABAB | 6 | XBAB | AXAB | ABXB | ABAX | -- |
| AXXB | 24 | -- | -- | -- | -- | -- |
| BABA | 3 | XABA | BXBA | BAXA | BABX | -- |

In the second pass two substitutions are performed simultaneously. Changing the two middle values of at-risk cell ABAB will allow them to be merged with the cell AXXB which is not at risk. Also note that cell AAAX’s count went from 13 to 16 because the 3 people who formerly had AAAA were combined with the 16 that have AAAX in the first pass.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Pass 2 (Double Substitution) | | | | | | | | |
| CELL | COUNT | xx-- | x-x- | x--x | -xx- | -x-x | --xx | SOLUTION |
| AAAX | 16 | -- | -- | -- | -- | -- | -- | -- |
| ABAB | 6 | XXAB | XBXB | XBAX | AXXB | AXAX | ABXX | **AXXB** |
| AXXB | 24 | -- | -- | -- | -- | -- | -- | -- |
| BABA | 3 | XXBA | XAXA | XABX | BXXA | BXBX | BAXX | -- |

The third pass performs three substitutions. This does help move BABA into a not-at-risk cell. No treatment is applied.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Pass 3 (Triple Substitution) | | | | | | |
| CELL | COUNT | -xxx | x-xx | xx-x | xxx- | SOLUTION |
| AAAX | 16 | -- | -- | -- | -- | -- |
| AXXB | 30 | -- | -- | -- | -- | -- |
| BABA | 3 | BXXX | XAXX | XXBX | XXXA | -- |

In the fourth and final pass, because the at-risk cell BABA hasn’t moved into a not-at-risk cell, the only solution is to remove all the demographic information of those 3 respondents. The combination of no demographic data and a work unit of at least 300 respondents greatly reduce their risk of being disclosed.

|  |  |  |
| --- | --- | --- |
| Pass 4 (Full Substitution) | | |
| CELL | COUNT | END SOLUTION |
| AAAX | 16 | AAAX |
| AXXB | 30 | AXXB |
| BABA | 3 | **XXXX** |