# 2021 Federal Employee Viewpoint Survey

# Public Release Data File

## Compressed File Contents:
- 2021_OPM_FEVS _PRDF.csv – Comma-delimited (CSV) Federal Employee Viewpoint Survey (FEVS) data file.
- 2021_OPM_FEVS _PRDF_Codebook.xlsx – Excel file containing information and context about the data.
- 2021_OPM_FEVS _PRDF Read Me.docx – Information about the CSV data file

## What's New in 2021?

Like the previous year, there were several changes to the 2021 OPM FEVS.

First, the survey itself was shorted by removing most of the items related to the COVID-19 Pandemic. The 38 core items from the previous survey remained. Please refer to the codebook for specific information regarding the items and their response choices.

Second, the survey was administered later in the year than usual – beginning of November and closing mid-December. This context is important to remember when examining any results in the survey. The previous survey was administered in the September-October timeframe, which itself was months later than the usual May/June administration in years prior.

Third, the survey in 2021 was a sample instead of a census, although most agencies ended up being a census anyway. More information can be found in the 2021 Technical Report when available.

As for changes to our Public Release Data File practices, nothing has changed since last year. The same demographic information appears, and the same process was used to mask sensitive data.

## Data File Summary

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

1. Core Survey Items: The "core" survey items which have appeared in previous surveys, appear in this year as well.

2. Modified Survey Items: A few items have had modifications since last year.
    a. 14-19; in 2020 these were broken out into three time periods, and have been condense to a single timeframe.
    b. 45 (minor edits to the text to reflect the new administration of the survey)
    c. 47-57

3. Demographics (all variables starting with "D"): This year's public use file includes nine demographics. Details of each can be found in the codebook.
    a. DRNO – Race/National Origin
    b. DHISP – Ethnicity
    c. DDIS – Disability Status
    d. DAGEGRP – Age Group
    e. DSUPER – Supervisory Status
    f. DFEDTEN – Federal Tenure
    g. DSEX – Sex
    h. DMIL – Veteran Status
    i. DLEAVING* – Intent to Leave, with modifications related to COVID

4. Work Unit Identifier: Used to identify where an employee works. For 2021, this is limited to only agency level, and only for agencies with a minimum respondent count of 750 or more.

5. Other: Includes the statistical weighting variable (POSTWT) and a randomly generated ID number (RandomID).

# Data Masking Methodology for Disclosure Avoidance

Starting in 2016, the FEVS 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 supervisory status categories into a more simplified Non-supervisor/Supervisor-type response.

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:

| Table 1: Example Demographic Profile | | |
| --- | --- | --- |
| **Demographic Category** | **Demographic Variable** | **Combined Group Profile** |
| SEX | (**B**) Female | |
| EDUCATION | (**B**) Bachelor's Degree | **B B X B** |
| 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 in the next section.

*For missing demographic data, a dummy value "X" is used.

# Masking Procedure Demonstration

In the first pass shown in table 2 three at-risk cells are identified (where count column is less than 10 and also shaded 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 no treatment is applied.

**Table 2:** 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 shown in table 3 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 13 that have AAAX in the first pass.

**Table 3:** 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 shown in table 4 performs three substitutions. This does not help move BABA into a not-at-risk cell. No treatment is applied.

**Table 4:** 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 shown in table 5, 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.

**Table 5:** Pass 4 (Full Substitution)

| CELL | COUNT | END SOLUTION |
|---|---|---|
| AAAX | 16 | AAAX |
| AXXB | 30 | AXXB |
| BABA | 3 | **XXXX** |