### Unique combinations of variables

[*The following is an example of an introduction paragraph*] All possible tables, from one-way tables to four-way tables, will be processed for factual (indirectly identifying) variables using the NCES *InitialRisk* software. Not all factual variables are needed as long as the selected ones are representative in terms of number of levels and distribution across the categories, and they include any that are in common with external files. The tabulations will be processed to identify categories of variables that are sparse when combined with other variables. By “sparse,” we refer to cell counts of less than three. This is related to identifying a sample unique, which is a set of values for a combination of variables that occurs only once in a sample. Only a few variables are necessary to be used in combination to result in a sample unique. Cell counts of two are also considered sparse since a sample member of the sparse cell of two knows its own data values and therefore can derive the other sample member’s data values. Combinations of variables that contribute most to the sparse cells are considered to have the highest risk, and therefore, the records that have rare combinations of variables are considered to have the highest risk. The analysis uses variables that may lead to a person being identified (information that is factual and publicly known or verifiable) such as demographic variables. Variables with subjective responses, such as cognitive items or personal attitudes, are not visible or identifiable by data intruders. Responses also can change on a daily basis. The subjective variables will be excluded from the extensive multi-way tabulations analysis.

For the analysis, continuous variables will be recoded into categorical variables and given a new name (R\_\*). This will be done by classifying each unit into deciles of the variable, where the deciles are derived across all units in the sample. Table X provides the list of factual variables that will be processed in *InitialRisk*.

Violations will be flagged for records that contributed to table cell counts that were less than three. The proportion of table cells with violations will be computed for each category of each variable. The algorithm counts the number of violations that involve a record for the set of tables generated. The process provides important information for determining the highest risk records in each data file.

Table X *InitialRisk* factual variables selected for analysis

| Level (e.g., school, student) | Variable Name | Label |
| --- | --- | --- |
|  | XXXX | XXXX’s label |

***InitialRisk* Processing**

For the purpose of the *InitialRisk* processing, and in order to help measure risk to inform the swapping process (especially with regards to selecting target records), several variables in the file have been recoded as follows:

[*List all recoded variables and their source variables, as well as how recoded variables are created.*]

***InitialRisk* Output**

Table 1 shows the risk measure (RM) resulting from the *InitialRisk* run. [*Briefly explain Table 1 here*.] Table 2 shows the variables and their categories that contributed to the highest percentage of violations.

Table [1]. [Survey name] Statistics for Violation Counts by Risk Stratum

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *InitialRisk*: Risk Stratum | N | Percent | Minimum | Median | Maximum | Mean | Sum |
| 0 |  |  |  |  |  |  |  |
| 1 |  |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |  |
| 3 |  |  |  |  |  |  |  |
| 4 |  |  |  |  |  |  |  |

[Resource: *InitialRisk* output]

Table [2]. [Survey name]: Percent Violations by Variable and Category

|  |  |  |  |
| --- | --- | --- | --- |
| Number of Variables Involved in Tables | Variable | Category of Variable | Proportion of cells with violations of Rule of 3 |
| 4 |  |  |  |
| 4 |  |  |  |
| 4 |  |  |  |
| 4 |  |  |  |
| 4 |  |  |  |
| 4 |  |  |  |

[Resource: *InitialRisk* output]

[*Given the initial risk assessment results, some indications of the recoding that will be performed should be mentioned in the DAP. Treatments such as, variable suppression, top-coding for continuous variables or ordinal variables with several categories, combining categories of categorical variables, should be indicated at least for key variables. NCES may also ask for a complete list of original variables and their treatments, and variables created through the SDC process.*]