### Data Swapping Plans

[*Here is an example set of paragraphs about the swapping process and one way to apply the risk measure from InitialRisk*. *The first paragraph is in the context of a school survey. Similarly the risks can be associated with a household survey and its PSUs, but typically to a lesser extent*.] The disclosure analyses focus on the two DRB-required components for the confidentialization of the [survey name] data. They include (1) the identification and masking of potential at-risk schools by comparing the study variables with publicly available data, using **probabilistic matching and deterministic swapping** on *school*-level data, and (2) the implementation of an additional measure of uncertainty of *school, student, and teacher* identification with the **random swapping** of data elements within the *school, student, and teacher* files. First we discuss the variable selection process.

### Selection of Variables for *DataSwap*

[As is always the case in identifying variables for *DataSwap*, one ideally wants to select variables that could be identifying but have limited analytical value whenever possible. In some cases demographic data should be selected and in others, responses to questionnaires may be preferable.

The process of variable selection is influenced by:

1. Number of files
2. Size of the files
3. Internal Linkages of the files
4. Linkages to external files
5. Data sources

*[Helpful hints:*

* *Assessment data normally not involved in swapping*
* *The number of variables selected from files for boundary and swapping affected by size of the files (number of records) – small files (e.g.) under 1000 records should only include a total of 3 or less variables. The variables should be limited to 4 categories or less.]*

### Swapping Specifications

Deterministic swapping can be done by assigning such designated cases to their own stratum with a sampling rate equal to 1. For the non-deterministic cases, the NCES *InitialRisk* software assigns a risk measure (RM) to each respondent. The RM assigns values 0 (low risk) to 5 (high risk) to records according to the number of table cell violations that the records were involved with. The RM will be used to give higher risk data records higher chance for swapping during the data swapping process. The actual swapping is conducted with the NCES *DataSwap* program, which ensures that a controlled random swapping method is applied in a manner consistent with IES DRB guidelines.

A benefit of swapping and using any other random perturbation method, is that there is added uncertainty by creating false positive matches that are indistinguishable from true positive matches by the data intruder. The intruder can match records with the known key characteristics or with an external file, but never knows if the result is a true match. Note that this type of uncertainty cannot be measured. The *DataSwap* approach to data swapping is a random perturbation method, and therefore it inherits this type of uncertainty for which the resulting risk level cannot be measured.

In the first step of *DataSwap*, target records will be selected via probability proportionate to size (PPS) sampling using the *RM* with higher chance of selection for higher risk data records. The proposed measure of size (*MOS*) is assigned to each record as follows: *MOS* = *RM2* + 1, where *RM* is the risk measure from *InitialRisk*.

Next, the values of the swapping variables will be exchanged between swapping targets and partners. The process reduces the disclosure risk, but at the same time incurs some bias. To reduce this bias, the swapping is set up to produce a higher probability for swapping for pairs with similar weights and/or close matches on specific characteristics; all records will be matched with each selected target case on characteristics that define swapping cells. The swapping cells are comprised of boundary variables (BOUNDARY) and swapping variables (SWAPVARS). The boundary variables are not allowed to be swapped. As discussed in the *DataSwap* manual, two potential swapping partners for the target record are initially selected, one from each neighboring (adjacent) cell. That is, within each neighboring cell, the record with the closest sampling weight to the target record is selected as a potential swapping partner. The search process continues by evaluating the swapping bias and the potential swapping partner. The record that results in the smallest swapping bias is chosen as the swapping partner.

To avoid the creation of logical inconsistencies, the swapping variables are linked to highly correlated variables in order to maintain internal consistency in the swapped records. Therefore, a list of swapping variables will be specified to ensure that data consistency will be retained with other closely-related variables. The parameter LINKSWAP will contain this variable, and will ensure that when one of the SWAPVAR variables changes value during the swapping process, the linked variable will change as well.

Several levels of controlled random swapping must be conducted. Each level can be done separately and independent due to the swapping variables being considered. The planned boundary and swapping variables are provided in a password protected separate file that serves as an appendix to the DAP.

At each level, we will process the swapping five times (using different random seeds) before selecting the best swapped dataset based on the *DataSwap* impact measures, and according to the IES DRB guidelines. The best run in terms of the global utility measures will be selected to serve as the official swapping file.

In order to preserve distributions by [X], the variable [X] will be used to help define the hard boundaries for determining the target case and its donor in the swapping procedure. Within these hard boundaries, the *DataSwap* program will select a donor case in the swapping cell closest to the case targeted for swapping. The weight variable used for the swapping process is the theoretical base weight for the relevant unit for the respective file.

To help check the impact of the swapping, a continuous variable, called KEYOUT, can be selected in *DataSwap* to evaluate the results on various utility measures computed in *DataSwap*, including correlations and regression coefficients. In addition, categorical variables, called KEYVARS, can be selected to evaluate the impact on measures of association. The KEYOUT and KEYVARS variables are provided in Table X. The swapping rates that are applied to each file in the *DataSwap* program will be chosen by the DRB chair. In addition, the swapping method is planned to be the “balanced” swapping method (SWAPMETH=2).

Table X Assignment of KEYOUT and KEYVARS variables

|  |  |  |
| --- | --- | --- |
| File | KEYOUT | KEYVARS |
| Survey Data Files | A list of KEYOUT variables | A list of KEYVARS |