NDACAN National Data Archive on Child Abuse and Neglect

Joining two data files from the RPG-2 dataset, using SAS.

This document is intended to assist data users with joining/merging two data files from RPG-2, dataset #218. The instructions are generally applicable for use with other data files with similar characteristics.

The two data files used in this example are the rpg_esl_person_level and the rpg_esl_case_level. These were selected because they feature the joining of two datasets where the unit of observation differs, resulting in different record counts for each file.

The rpg_esl_person_level dataset contains information about each person who participated in the RPG-2 data collection, including substance using adults and their children. As shown in the User's Guide Table 1, you will note that the variables that identify a unique record for the rpg_esl_person_level data file are: Grantee_ID, Case_ID, and Ind_ID. In this file, because each case has multiple people and the unit of observation is the person, the Case_ID repeats across records for each person represented in that case. The variables that identify a unique record in the rpg_esl_case_level data file are: Grantee_ID and Case_ID. Since Grantee_ID and Case_ID are unique to both data files, we will merge/join the records based on those unique identifiers.

In this example, we are interested in preserving all records from the rpg_esl_person_level data file that contains 8,626 records pertaining to each individual who participated in the RPG-2 study. Due to the rpg_esl_case_level data file containing information about the "case," there are only 2,887 records. To summarize, each **case** represents multiple **individuals** (substance using adults and their children). When we merge/join these two datasets together we want the information about the case properly associated with records for the individuals. This is known as a many-to-one relationship. The Proc SQL procedure with the Left Join specification will create a joined file consisting of all records from the rpg_esl_person_level and the matching records from the rpg_esl_case_level file.

The instructions are broken out into sections and should be followed in the order in which they appear. Data users should copy each section of code into the Editor window of SAS, omitting section headings and other explanatory text. You will need to update all file paths to point to the location of the data file on your own computer and provide locations to where the modified data files should be saved.

The instructions were written based on Version 9.4 (64-bit) of the SAS software package.

Section One: Sort the rpg_esl_person_level data file

This section will assist data users with sorting the rpg_esl_person_level data file. The Proc SQL procedure with the Left Join specification does not require that each of the input data files are sorted prior to the join, however, if the input files are not sorted before the join, then they will not be sorted in the resulting joined data file. We have chosen to sort both files first.

```
LIBNAME library1 “c:\temp\change-the-file-path-location-to-
point-to-folder-with-rpg-esl-SAS-data”;

RUN;

PROC SORT DATA=library1.rpg_esl_person_level OUT
=library1.rpg_esl_person_level_sorted;

 BY grantee_id case_id;

RUN;
```

Section Two: Sort the rpg_esl_case_level data file

This section will assist data users with sorting the rpg_esl_case_level data file prior to conducting the Proc SQL left join procedure in Section Three.

```
PROC SORT DATA=library1.rpg_esl_case_level OUT
=library1.rpg_esl_case_level_sorted;

 BY grantee_id case_id;

RUN;
```

Section Three: Merge/join the rpg_esl_person_level data file with the rpg_esl_case_level data file.

In this section, you will perform the data join/merge of the rpg_esl_person_level data with the rpg_esl_case_level data file. The SAS code below will store the joined file in the same location as the source data files used in the join.

```
PROC SQL;

CREATE TABLE library1.merged_person_and_case_level AS

     SELECT *

     FROM library1.rpg_esl_person_level_sorted LEFT JOIN
     library1.rpg_esl_case_level_sorted

     ON
     rpg_esl_person_level_sorted.Grantee_ID=rpg_esl_case_level_s
     orted.Grantee_ID AND rpg_esl_person_level_sorted.CASE_ID=
     rpg_esl_case_level_sorted.CASE_ID;
```

```
QUIT;
```

In this example, you should have a combined data file that contains 8,626 records and 67 variables. You should refer to each respective data file's codebook for information pertaining to the variables in the newly joined/merged dataset.

This document was prepared by NDACAN Analyst, Holly M. Larrabee. Email questions to NDACANsupport@cornell.edu.