# Lab Exercise #4: Validating Links

## Learning Goals

- Consider linking validation criteria
- Write validation code for both long and/or wide files

## Summary

In this exercise, you will link validate links created using CPSIDP in previously completed exercises. Validation will be based on demographic variables (AGE, SEX, and RACE). Recall that the CPS is a survey of individuals who live in physical structures, and the same physical location is in the CPS for the entire rotation, even if the individuals who live in the structure change over time. While original CPS identifiers *should* uniquely identify individuals over time, they do not always. To guard against false matches, we recommend double checking links using AGE, SEX, and RACE. For the purposes of this exercise, assume that RACE and SEX should remain constant across time for individuals in the CPS. AGE, however, should change in expected ways; specifically, it should increase over time. For the purposes of this exercise, any links that do not match on these demographic variables across all observations in the linked file will be considered invalid.

## Part 1: Validation Criteria

1. How might you approach validating links between multiple time points in a data file?

   _____
   _____
   _____
   _____
   _____

2. What would a reasonable amount of change to tolerate in AGE before calling a link invalid for a month-to-month link? For a link made between the same month in adjacent years? Look at AGE codes – what should the rule be if you're looking at the oldest CPS respondents?

   _____
   _____
   _____
   _____
   _____

3. In addition to simply comparing the AGE, SEX, and RACE values in the data, what other pieces of information might you want to draw on to construct your rules?

   _____
   _____
   _____
   _____
   _____

## Part 2: Applying Validation Criteria

4. Return to *part 1 (Month-to-Month Linking) of exercise 1 on linking possibilities*. Using a long file, validate the links between the August and September 2015 samples using AGE, SEX, and RACE. Fill in Table 1 with the number of validated links made for each August/September month-in-sample pair.

**Table 1**

| | September (Volunteer) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| August (Veterans) | *MISH* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* |
| | *1* | | | | | | | | |
| | *2* | | | | | | | | |
| | *3* | | | | | | | | |
| | *4* | | | | | | | | |
| | *5* | | | | | | | | |
| | *6* | | | | | | | | |
| | *7* | | | | | | | | |
| | *8* | | | | | | | | |

5. Explain what you see in the August MIS 4 and 8 rows and the September MIS 1 and 5 columns.

   _____

   _____

   _____

6. What are the allowable age differences in this linked file? _____
7. What proportion of records in both the 2015 Veterans and Volunteer supplements are valid based on AGE? _____ SEX? _____ RACE? _____ All three characteristics? _____
8. *Reference part 2 (Full Panel Linking) of exercise 1 on linking possibilities*. Return to the linked file you created containing all 8 observations of those entering the CPS in August 2015. What proportion of those linked records is valid across all eight observations based on AGE? _____ SEX? _____ RACE? _____ All three characteristics? _____
9. Return to *exercise 3 on linking the Food Security supplement to the ASEC*. Using a wide linked file, validate the links made in this exercise using AGE, SEX, and RACE. How many records link mechanically and are valid based on AGE? _____ SEX? _____ RACE? _____ All three characteristics? _____

Answers

## Part 1: Validation Criteria

1. How might you approach validating links between multiple time points in a data file?
   There are many ways to approach validation. The IPUMS CPS approach to linking validation looks for exact matches on SEX and RACE for each individual's full set of observations. AGE may either remain constant or increase by one year within MIS 1-4 or MIS 5-8 and may increase by up to two years when going between MIS 1-4 and MIS 5-8. Note that the AGE topcode must be addressed (see answer to #2 below).
2. What would a reasonable amount of change to tolerate in AGE before calling a link invalid for a month-to-month link? For a link made between the same month in adjacent years? Look at AGE codes – what should the rule be if you're looking at the oldest CPS respondents?
   If a link is between adjacent months, then an increase of one year would be allowable. If the linked file contains records from two different years, then up to two years would be a reasonable allowable increase. The top code in AGE must also be accounted for. For those age 80 when first observed, five years would be an allowable increase if the linkage did or did not span multiple years. For those age 79 when first observed, one year is the only allowable increase because the AGE code of 80 represents those who are 80-84 years old.
3. In addition to simply comparing the AGE, SEX, and RACE values in the data, what other pieces of information might you want to draw on to construct your rules?
   You may want to include other variables that change in expected ways. For example, educational attainment should stay the same or increase.

## Part 2: Applying Validation Criteria

4. *Return to part 1 (Month-to-Month Linking) of exercise 1 on linking possibilities*. Using a long file, validate the links between the August and September 2015 samples using AGE, SEX, and RACE. Fill in Table 1 with the number of validated links made for each August/September month-in-sample pair.

**Table 1**

| | September (Volunteer) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| August (Veterans) | *MISH* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* |
| | *1* | 0 | 14,589 | 0 | 0 | 0 | 0 | 0 | 0 |
| | *2* | 0 | 0 | 15,258 | 0 | 0 | 0 | 0 | 0 |
| | *3* | 0 | 0 | 0 | 15,331 | 0 | 0 | 0 | 0 |
| | *4* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | *5* | 0 | 0 | 0 | 0 | 0 | 15,034 | 0 | 0 |
| | *6* | 0 | 0 | 0 | 0 | 0 | 0 | 15,553 | 0 |
| | *7* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 15,220 |
| | *8* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

5. Explain what you see in the August MIS 4 and 8 rows and the September MIS 1 and 5 columns. Those in MISH 4 in August, begin their 8-month break in September. Those who are in MISH 8 in August are out of the sample for good in September.
6. What are the allowable age differences in this linked file? 1 and 5 years (due to the top code)
7. What proportion of records in both the 2015 Veterans and Volunteer supplements are valid based on AGE? 91,095 SEX? 91,551 RACE? 91,540 All three characteristics? 90,985
8. *Reference part 2 (Full Panel Linking) of exercise 1 on linking possibilities*. Return to the linked file you created containing all 8 observations of those entering the CPS in August 2015. What proportion of those linked records is valid across all eight observations based on AGE? 9,622 SEX? 9,861 RACE? 9,861 All three characteristics? 9,548
9. *Return to exercise 3 on linking the Food Security supplement to the ASEC*. Using a wide linked file, validate the links made in this exercise using AGE, SEX, and RACE. How many records link mechanically and are valid based on AGE? 27,636 SEX? 28,076 RACE? 28,075 All three characteristics? 27,520