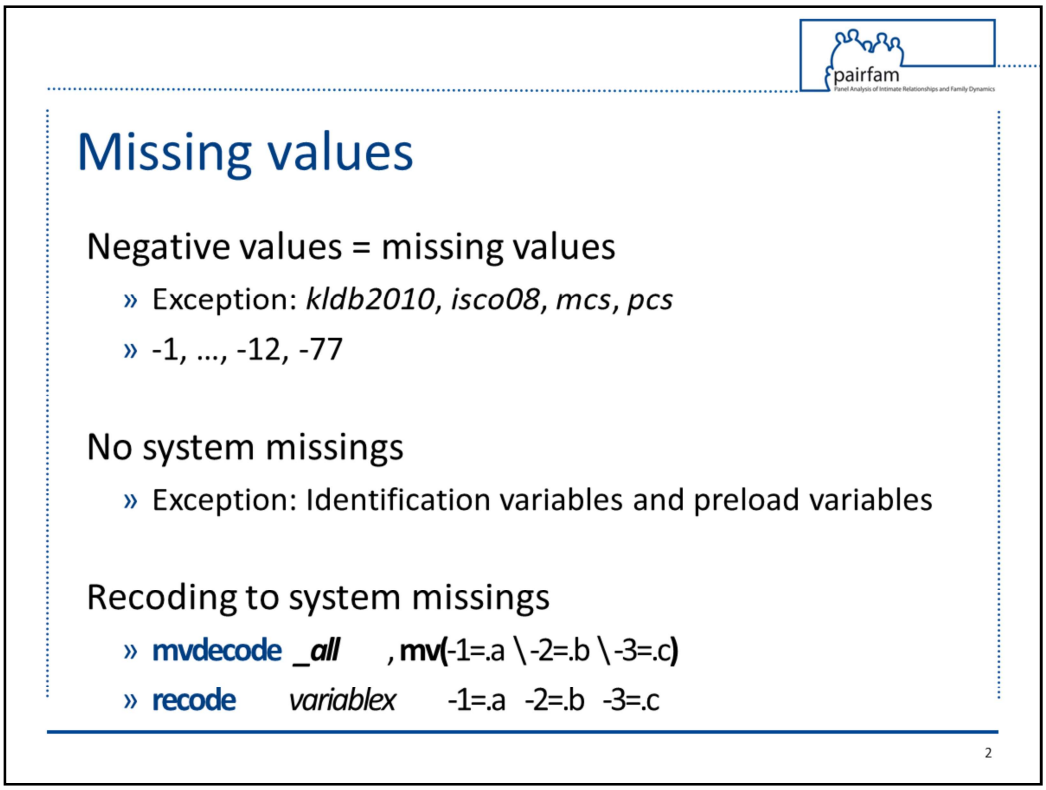
The corresponding video tutorials are available online:
https://www.youtube.com/playlist?list=PL7BcpOtSe5u_zQctYXz4ee79Zc9r4mfnr

pairfam tutorial

8. Missing values and filter missings

Kristin Hajek & Madison Garrett, May 2022

The following tutorial explains the usage of missing values and filter missings.



How does *pairfam* manage missing values?

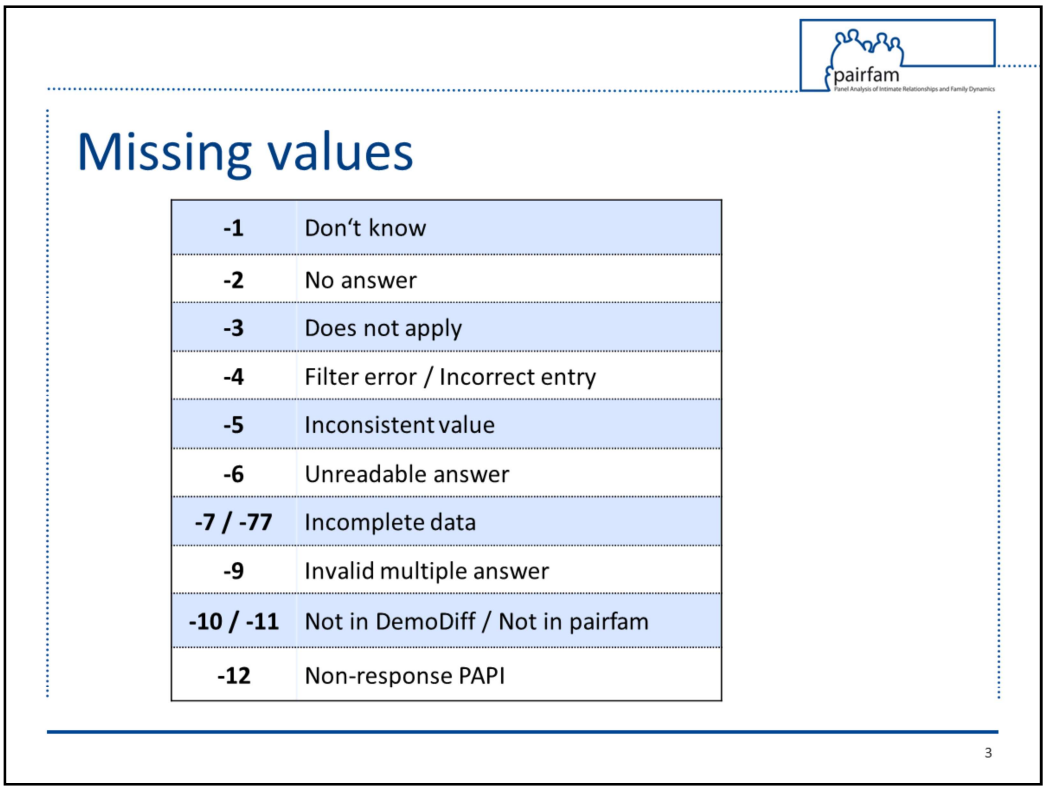
Missing values in the *pairfam* data are typically coded with negative numbers.

Only four generated variables have valid negative values: *kldb2010*, *isco08*, *mcs*, and *pcs*.

Missing values range from -1 to -12. Additionally, *mcs* and *pcs* also contain the missing code -77.

With the exception of identification variables (such as *id* or *pid*) and preload variables, the data include no system missings.

Negative missing values can be easily recoded into system missings with the Stata command *mvdecode* or *recode*. Alternatively, analyses can also be restricted to positive values only.



Why are there so many different missing values?

The missing values -1 “Don’t know” and -2 “No answer” were assigned if the respondent could not or did not want to answer a question. These codes are the only missing values documented in the questionnaire.

The value -3 “Does not apply” was assigned if a respondent had not been posed the corresponding question, meaning the respondent was “filtered over”.

Errors in the CAPI program that erroneously guided respondents to the wrong questions in the interview are indicated by the missing code -4 “Filter error / Incorrect entry”, as are incorrect data entries made by interviewers. In waves 12 and 13, the value -4 was also assigned if the respondent incorrectly responded (or incorrectly did not respond) to a question in PAPI mode.

In order to detect inconsistencies between responses, we checked for logically impossible or empirically implausible combinations of values for two or more variables. Inconsistent values were then coded to -5 “Inconsistent value” if it was clear that the value was implausible.

For open answers that were not legible, the value -6 “Unreadable answer” was assigned.

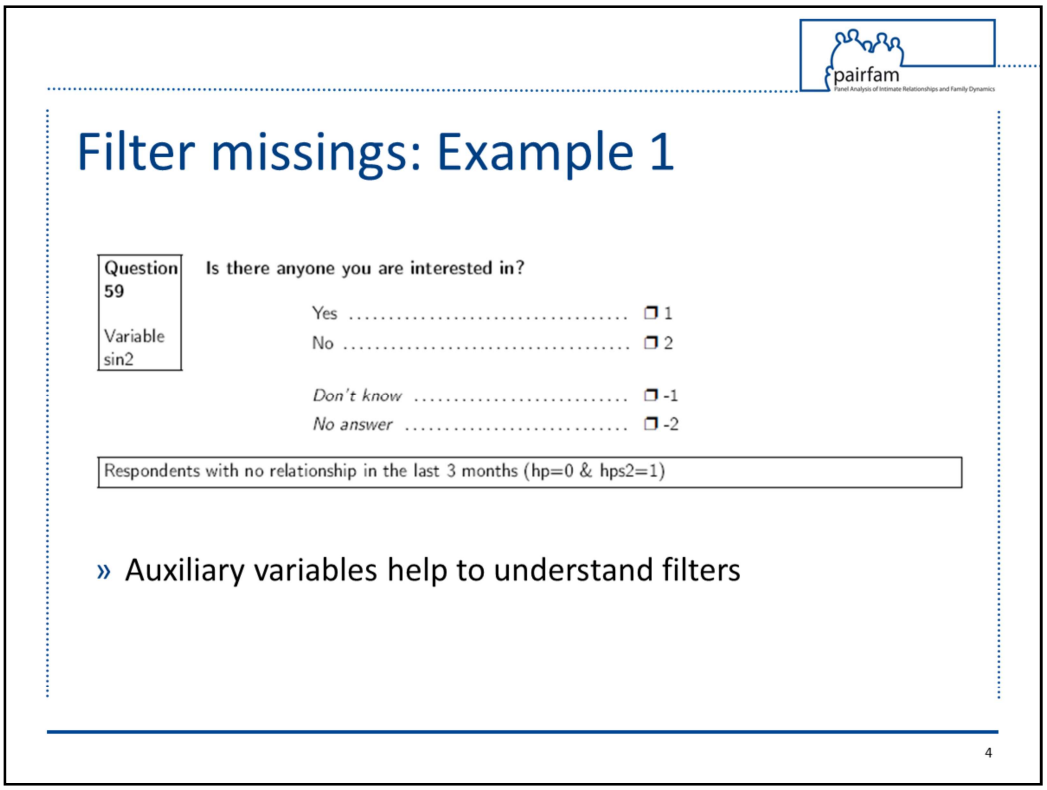
For generated data variables and files, the value -7 “Incomplete data” is used to indicate cases for which the necessary information was lacking to compute a valid value. For the generated variables *mcs* and *pcs,* the value -77 “Incomplete data” is

used.

In the alteri data sets, the value -9 “Invalid multiple answer” was assigned if the respondent checked more than the permitted number of options.
For waves 1 and 3, special missing codes indicate differences between the *pairfam* questionnaire and the *DemoDiff* sample. If a question from the *pairfam* questionnaire was not part of *DemoDiff*, the corresponding variable was set to -10 “Not in *DemoDiff*”. *DemoDiff* variables not included in the *pairfam* questionnaire are indicated by -11 “Not in *pairfam*”.

In waves 12 and 13, anchor respondents interviewed via telephone due to COVID-19 restrictions who did not return the PAPI questionnaire received the missing value -12 “Non-response PAPI” for these questions.

Please be aware that the values -10, -11, and -12 are also part of the classification scheme for *kldb2010* and *isco08*.



What exactly are filter missings?

Let's look at an example. Single respondents were asked the question "Is there anyone you are interested in?".

This is a screenshot from the anchor codebook. On the left you can see the question number and the corresponding variable name.

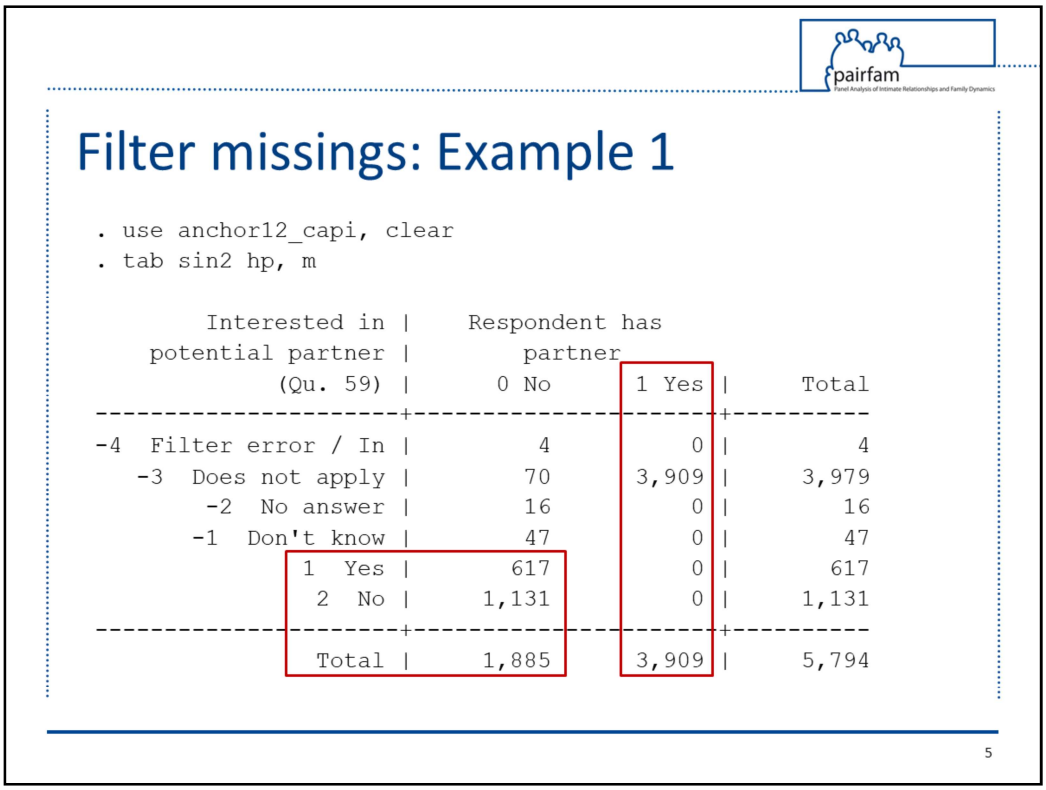
Here you can see the question wording and the available answers as well as their variable values.

If a filter was implemented for a question, there is a box underneath the question stating which respondents were posed the question of interest – in text and numerical form. The variables in parentheses are the ones used for filtering.

In this example, this question was posed only to respondents with no reported relationship in the last three months.

As you can see, auxiliary variables, such as *hp* and *hps2*, can help you to understand question filters and facilitate your analyses.

It is helpful to consult the codebooks and check question filtering to better understand a variable's values.

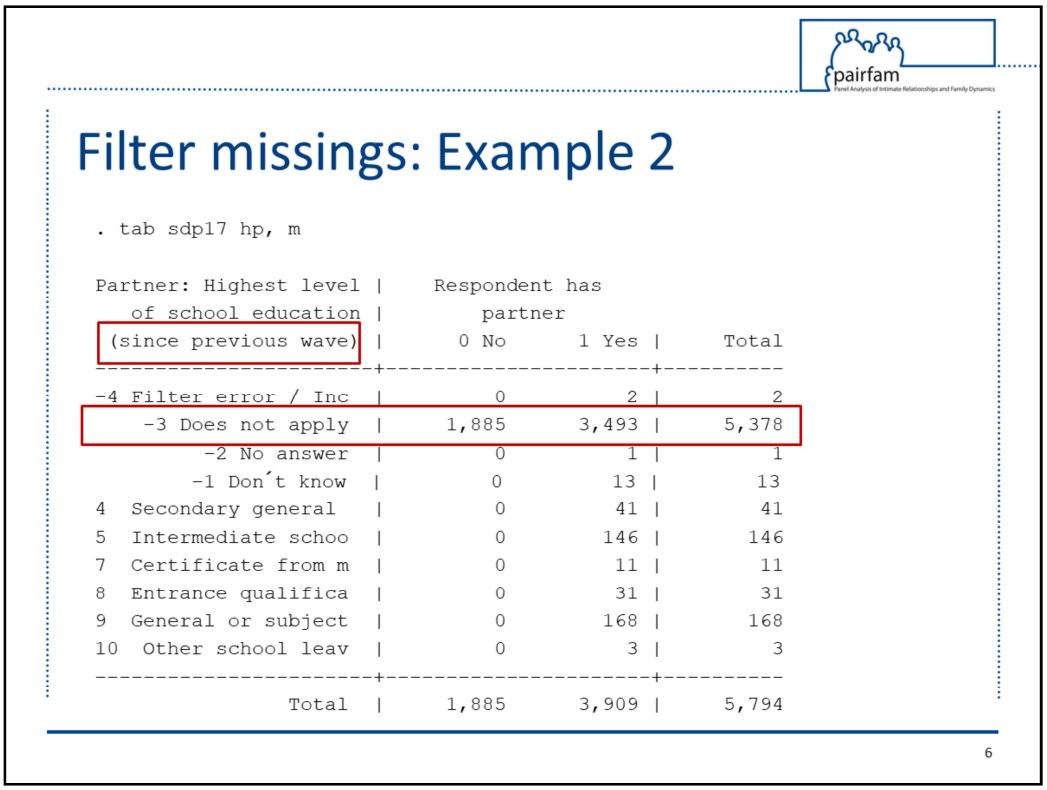


To further this example, let's look at the anchor12_capi data set in Stata.

Here I have tabulated the variables *sin2* and *hp*.

Respondents who have a partner receive the missing code -3 "Does not apply" for *sin2* as they were not posed this question.

Only a small portion of participants, the ones who were single at the time of the interview for wave 12, answered this question.



The next example illustrates why consulting the codebooks to understand the filtering of certain questions can be important, and how generated variables can be useful for your analyses.

Here, I tabulated the partner's highest level of school education – variable *sdp17*.

Naturally, anchor respondents without a partner all have the missing code -3 "Does not apply".

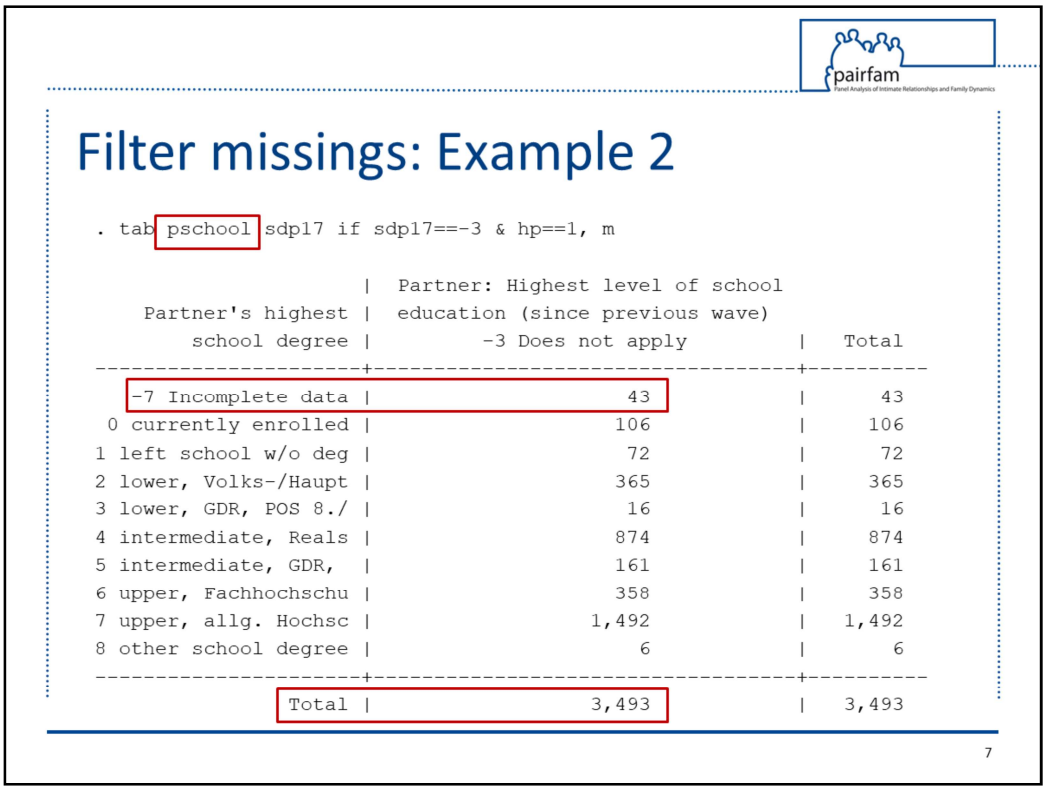
However, there are also almost 3,500 respondents who do have a partner with this missing value as well.

If you read the question closely, you can see that the variable *sdp17* asks for the highest level of school education since the previous wave.

The filter is defined as follows: "Respondents with partner from previous wave who received a school leaving certificate since the previous wave, or with new partner who received a school leaving certificate".

As the goal is not to overburden respondents, they are asked to only report their partner's highest school education ***if*** there has been a change since the previous wave.

Does this mean that you will have to check each wave of the anchor data and copy the highest level of school education to all the other waves? No!

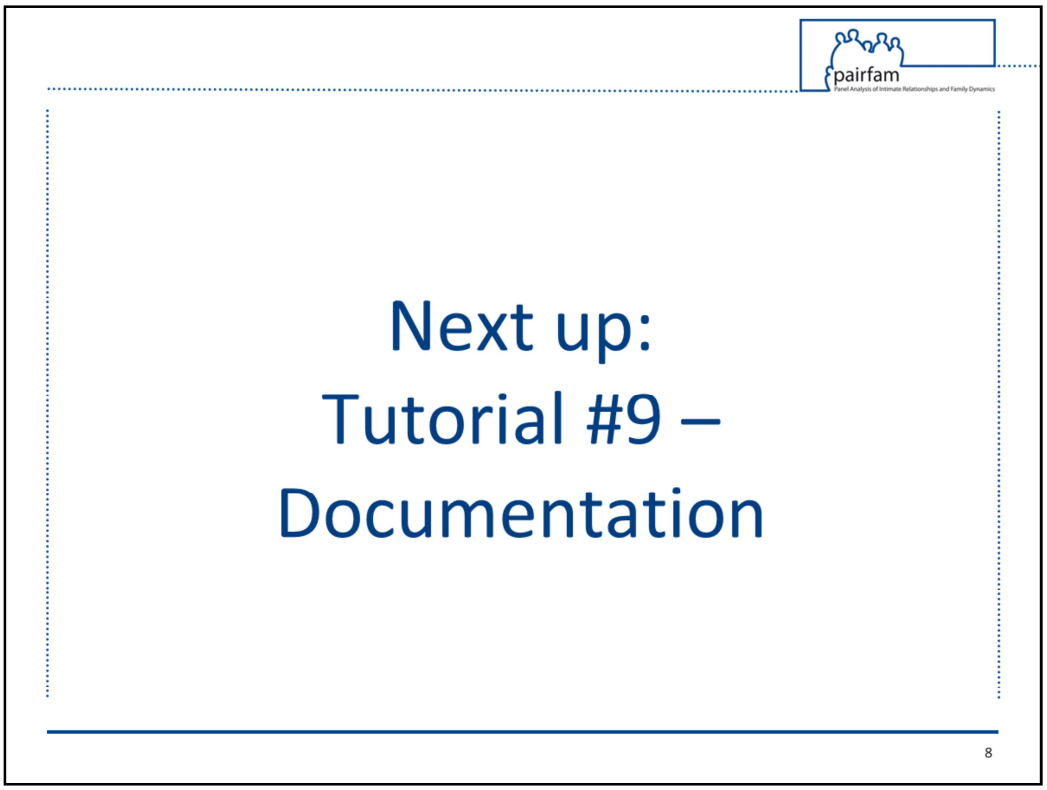


Because we want to make data analysis as easy as possible for our users, the generated variable *pschool* sums up all information on the partner's highest school degree across all waves.

Here, I tabulated *pschoo*l for the almost 3,500 respondents who had the missing value -3 for *sdp17*.

As you can see, only 43 respondents still have missing values while the rest have valid codes.

We highly recommend consulting the generated variables delivered in the data, as they really can facilitate data analysis!



We’ve reached the end of the eighth tutorial.

Tutorial number 9 will present the available documentation included in the Scientific Use File.