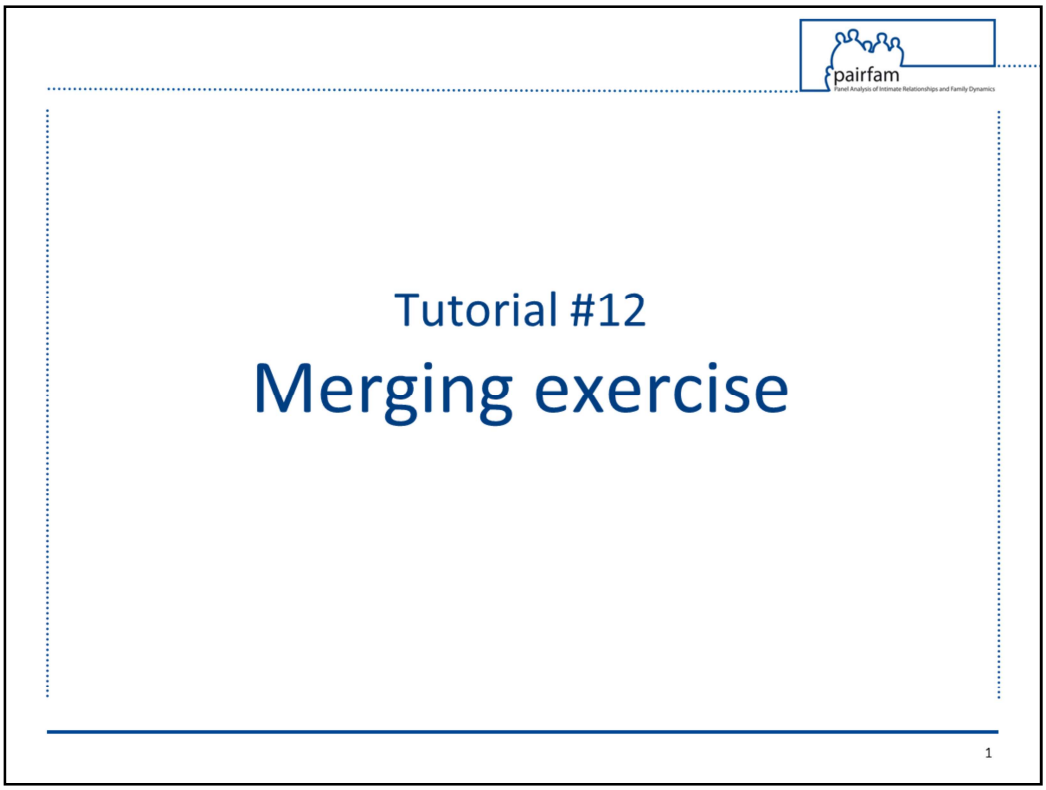
The corresponding video tutorials are available online:
https://www.youtube.com/playlist?list=PL7BcpOtSe5u_zQctYXz4ee79Zc9r4mfnr

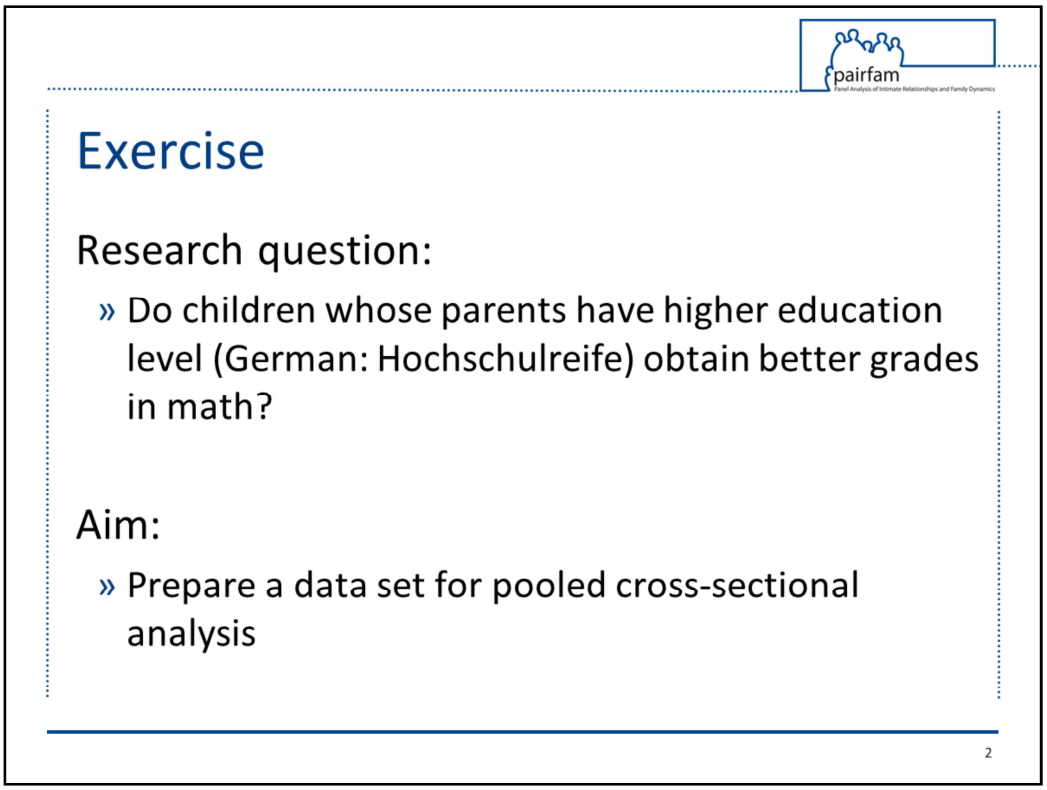


pairfam tutorial

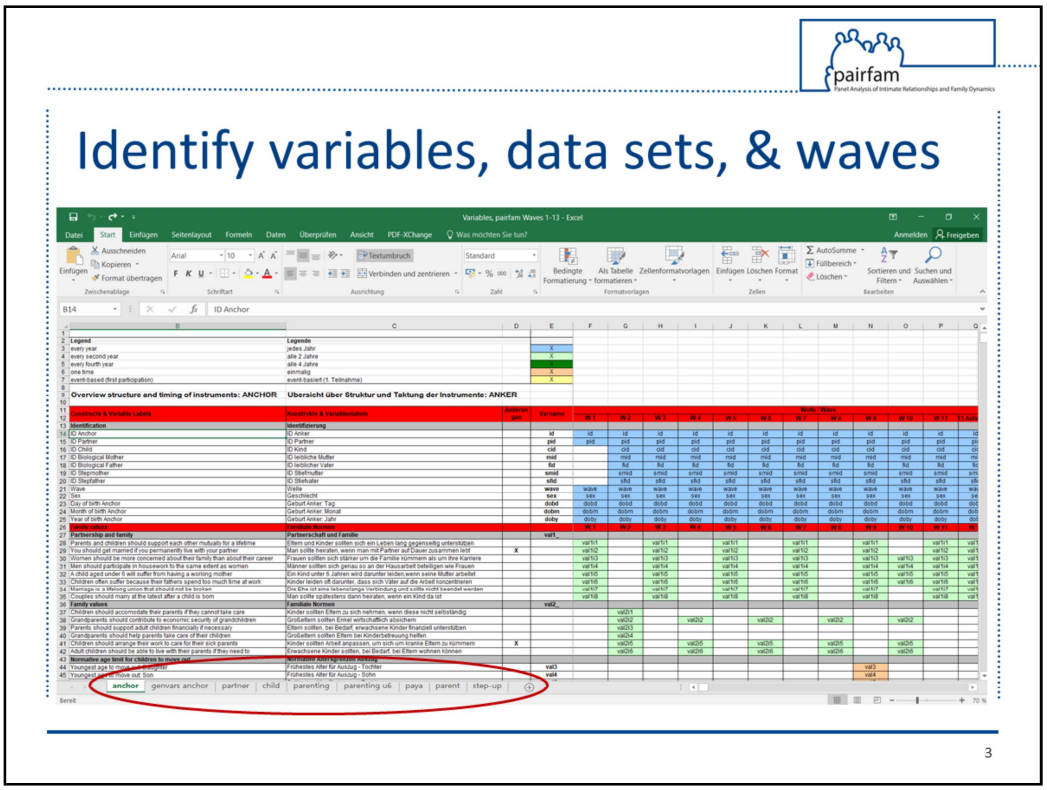
12. Merging exercise

Kristin Hajek & Madison Garrett, May 2022

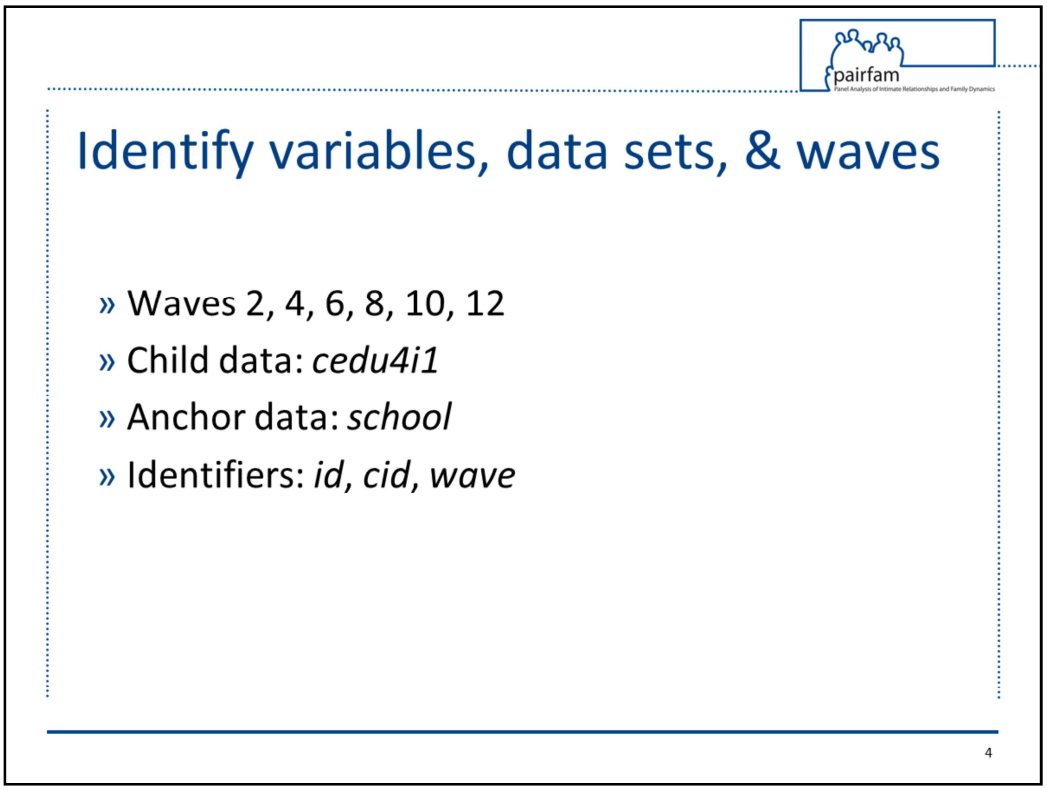
This tutorial represents a hands-on exercise for merging *pairfam* data in Stata. You will be guided through the process of matching the *pairfam* data sets based on a given research question in order to prepare a data set that enables you to analyze this question.



The research question is: Do children whose parents have higher education level (in German: Hochschulreife) obtain better grades in math? The aim of the matching process is to prepare a data set for pooled cross-sectional analysis. This means, that these questions might have been asked in several waves and I want you to merge data from several waves in long format. Take a moment to write down this question and think about where you would start. Please pause the video tutorial now and think about your next step.



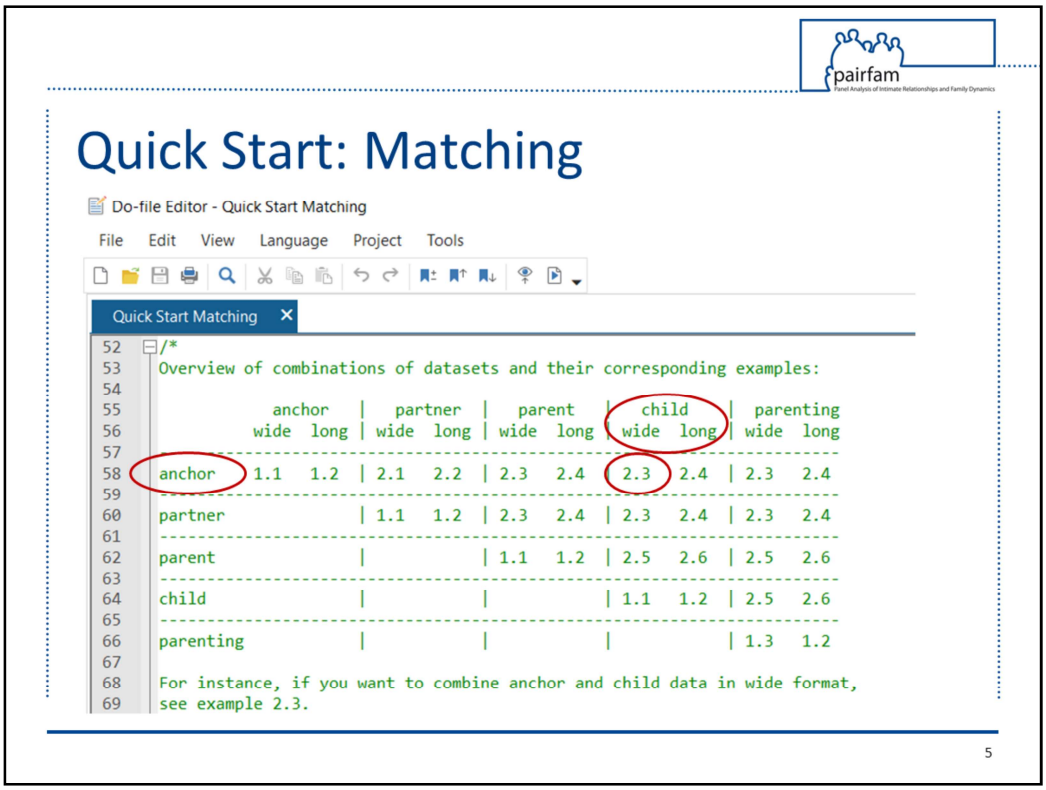
As you will remember, the first step in the matching process is to identify the variables, data sets, and waves in which relevant variables are stored. You might find the variables for answering this research question in the parent and anchor data sets, or in the anchor and child data sets. Please be aware that the variables for the different respondent types are stored on separate Excel sheets. Pause the video tutorial now to open the Question Program and search for the appropriate variables.



Have you found the appropriate variables?

Children‘s grades in math are included in the child data sets for waves 2, 4, 6, 8, 10, and 12 with the variable *cedu4i1*. In the anchor data sets, you can find the variables *sd32i1-13,* which store the highest educational level the anchor respondent has attained since the previous wave. However, it is easier to use the generated variable *school,* pre-prepared by the *pairfam* team to simplify your research. As you may remember, the generated variables are also stored on a separate sheet in the Question Program.

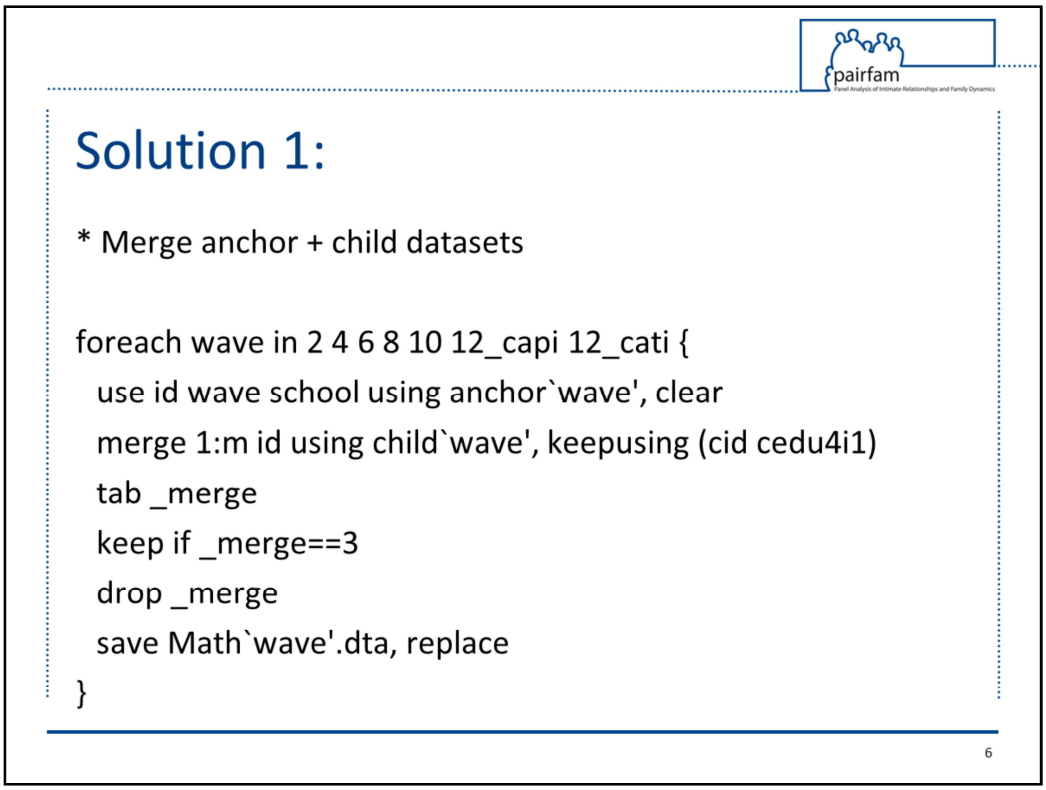
Before we start with the matching process, we must consider which identifiers are needed to match the anchor and child data from waves 2, 4, 6, 8, 10, and 12. As you already know, each data set contains the anchor identifier *id* and the time variable *wave*. The children identifier *cid* might also be useful to identify the anchor‘s children if more than one child has been interviewed. However, *cid* is not absolutely necessary for the matching process.



Now that we have identified all relevant variables and data sets, you can open the Quick Start "Matching", which will assist you in the matching process.

At the top of the Quick Start you will find an overview of combinations of data sets and corresponding examples. For instance, if you want to combine the anchor and child data in wide format, you will find an example in Section 2.3.

Pause this tutorial now and try to match the anchor and child data from waves 2, 4, 6, 8, 10, and 12 with the help of the Quick Start "Matching".

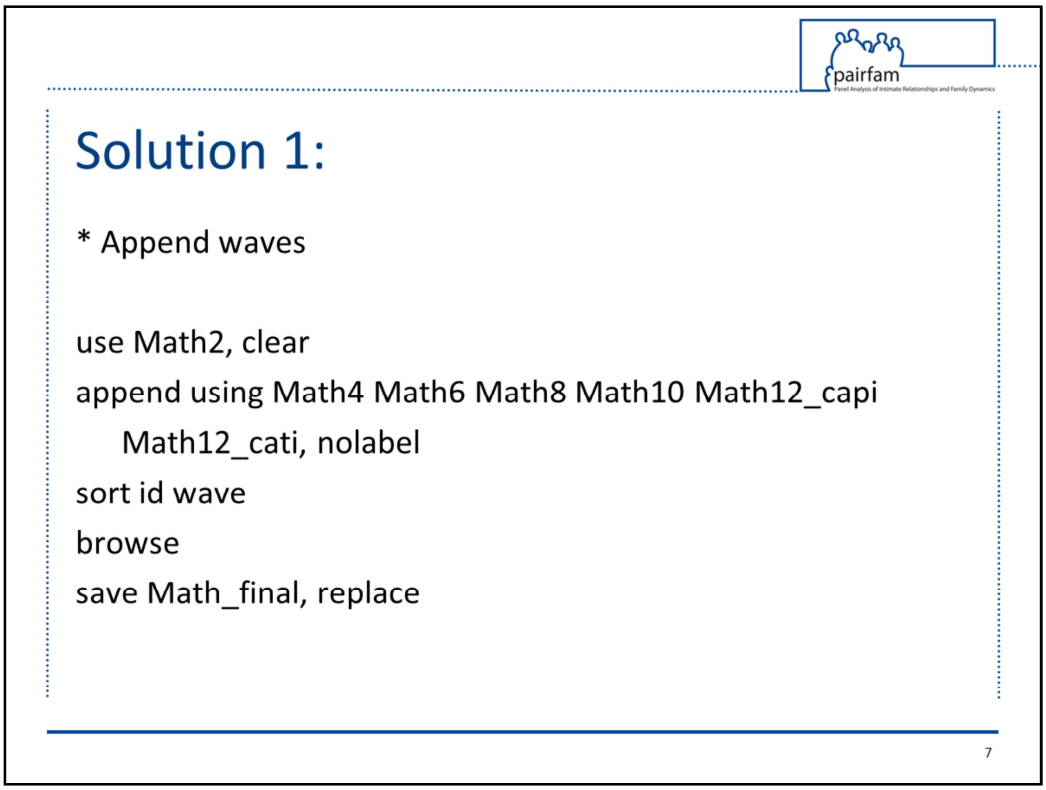


There are several ways to match the data sets, depending on your starting point. We will cover two different approaches.

After you have defined the data path (where the data is stored on your computer), you can begin. One approach to matching pairfam data sets is to always start with the anchor data.

The first step is to merge the anchor with the child data sets. One option is to define a loop to run the following for waves 2, 4 6, 8, 10, and 12. Please note that the wave 12 data is stored in two data sets, one for CAPI respondents and one for CATI respondents.

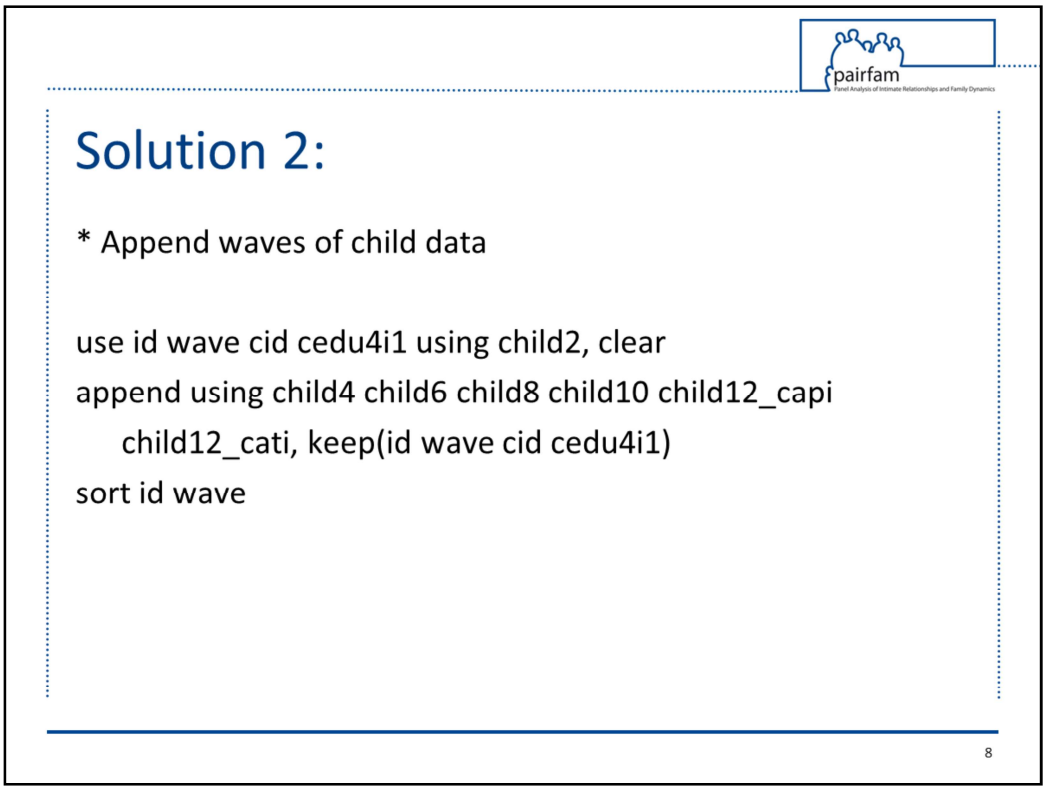
To do this, open the varibles *id*, *wave,* and *school* from the wave 2 anchor data with the command *use id wave school using anchor`wave', clear*. Then, merge one anchor respondent to many children using the child data from wave 2, keeping only the variables *cid* and *cedu4i1*. Afterwards, tabulate the automatically generated variable *_merge* and keep only the cases that were able to be matched. Anchor respondents without any child interviewed are therefore excluded from the analytical sample. The variable *_merge* can then also be dropped, and the merged data saved as *Math2*. This same process will loop automatically over waves 4 to 12.



Now all we need to do is append all the waves of merged data.

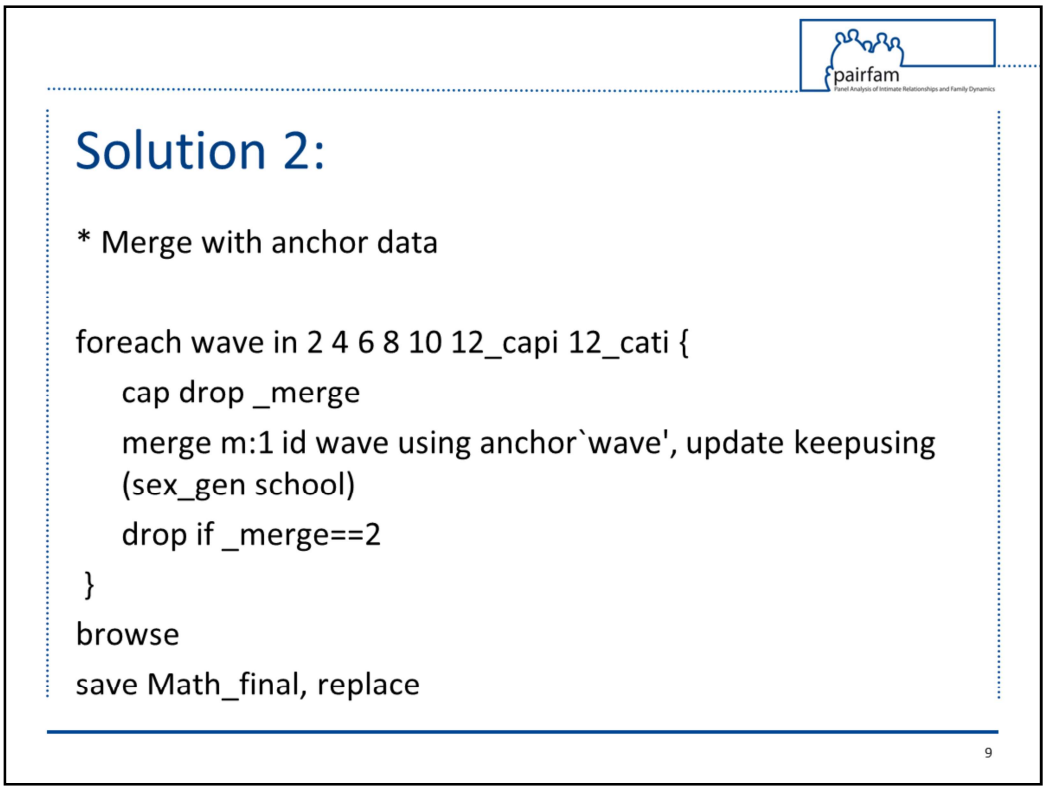
Open the data set *Math2* and append the data sets for waves 4, 6, 8, 10, and 12.

Afterwards, sort the data and check the browser to see whether the matching process was successful, then save the final data set.



The second approach is to start with the append command and the child data sets.

Load the variables *id, wave, cid,* and *cedu4i1* from the *child2* data set. Then, append the same variables from the child data from waves 4, 6, 8, 10, and 12.



Now is when things get tricky. Be careful when merging the anchor data to the newly appended child data set.

Again, one option is to define a loop to run over waves 2, 4 6, 8, 10, and 12.

First, exclude the variable *_merge,* if it exists. Then, merge many children to one anchor – this time we need two identification variables (*id* and *wave*) because we have already appended all waves of the child data. Furthermore, adding the command "update" is very important here; if you forget it, Stata might overwrite cases you have previously merged.
As there are some pitfalls with this approach, we recommend using the first. However, this approach is also viable and might be more intuitive for some users.

After you have matched the data sets, you can begin the data preparation process and, after, your analysis.

This marks the end of the *pairfam* tutorials. Thank you for listening and good luck with your research!