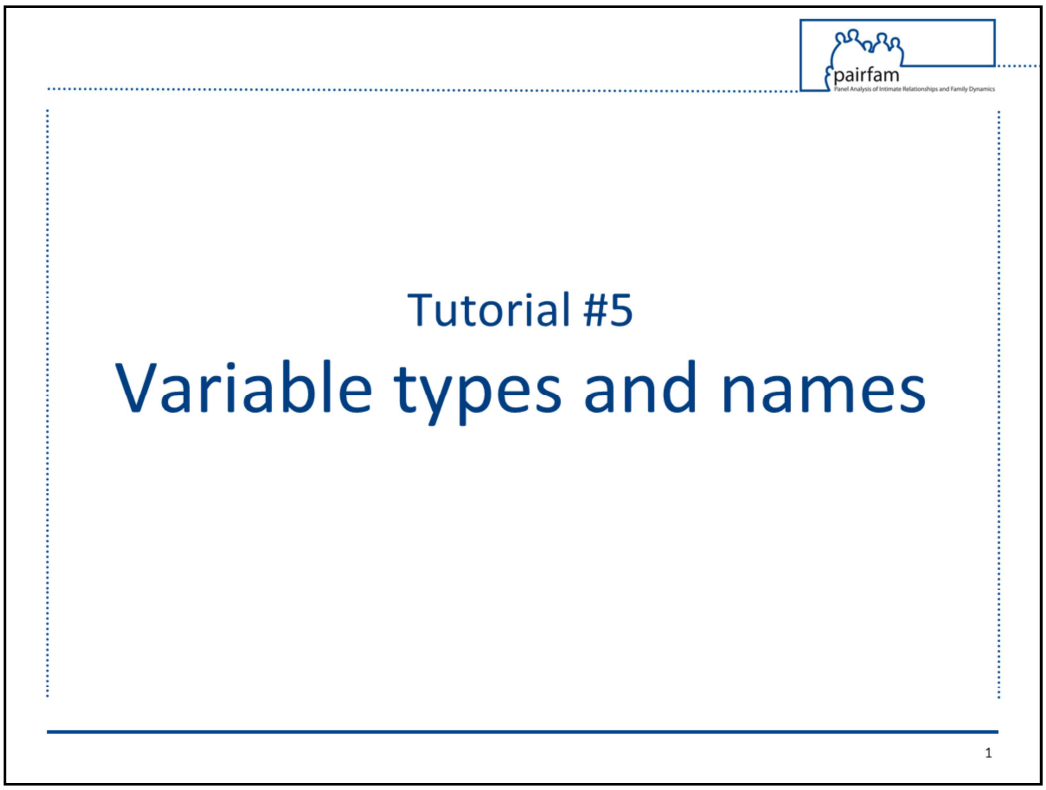
The corresponding video tutorials are available online:
https://www.youtube.com/playlist?list=PL7BcpOtSe5u_zQctYXz4ee79Zc9r4mfnr

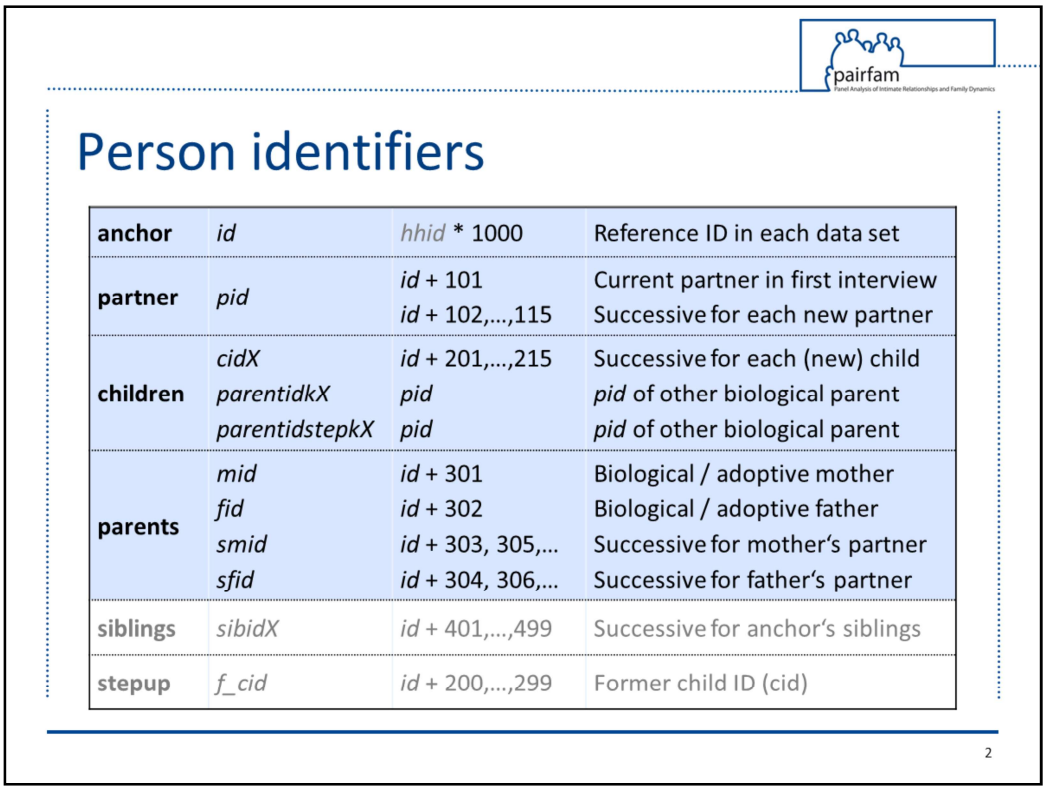


pairfam tutorial

5. Variable types and names

Kristin Hajek & Madison Garrett, May 2022

This fifth tutorial covers all variable types included in the *pairfam* main data sets and explains their naming convention.



Person identifiers

anchor	*id*	*hhid* * 1000	Reference ID in each data set
partner	*pid*	*id* + 101 *id* + 102,...,115	Current partner in first interview Successive for each new partner
children	*cidX* *parentidkX* *parentidstepkX*	*id* + 201,...,215 *pid* *pid*	Successive for each (new) child *pid* of other biological parent *pid* of other biological parent
parents	*mid* *fid* *smid* *sfid*	*id* + 301 *id* + 302 *id* + 303, 305,... *id* + 304, 306,...	Biological / adoptive mother Biological / adoptive father Successive for mother's partner Successive for father's partner
siblings	*sibidX*	*id* + 401,...,499	Successive for anchor's siblings
stepup	*f_cid*	*id* + 200,...,299	Former child ID (cid)

2

What are "person identifiers"?

The "identifiers" are identification variables for the anchor and alteri respondents. Each respondent was assigned a unique identifier in their first wave of participation that remained unchanged throughout the panel. The variable *hhid* is the 3-6 digit household number assigned by Kantar Public to identify anchor respondents in each wave.

The variable *id* was then generated by multiplying *hhid* by 1000, and serves as the reference identifier for anchor respondents. This variable can be found in every data set, so that users can easily match anchor data to alteri data.

The partner identification variable *pid* is generated by adding 101 to the anchor's *id* for the current partner in the first interview – for example, the anchor with the *id* 999000 has a partner with the *pid* 999101. New partners were numbered consecutively; continuing with the same example, the same anchor's second reported partner would have the *pid* 999102.

From wave 2 onward, children are identified by the variable *cid.* Each child mentioned by the anchor respondent was assigned a number at first mention.

Similar to the partner's *pid*, *cid* is generated by adding 200 to the anchor respondent's *id,* plus the position number of the child in the anchor data – for example, 999202 for the second child who was mentionend in the anchor interview.

The variable *parentidk* contains the identification number of a child's other biological

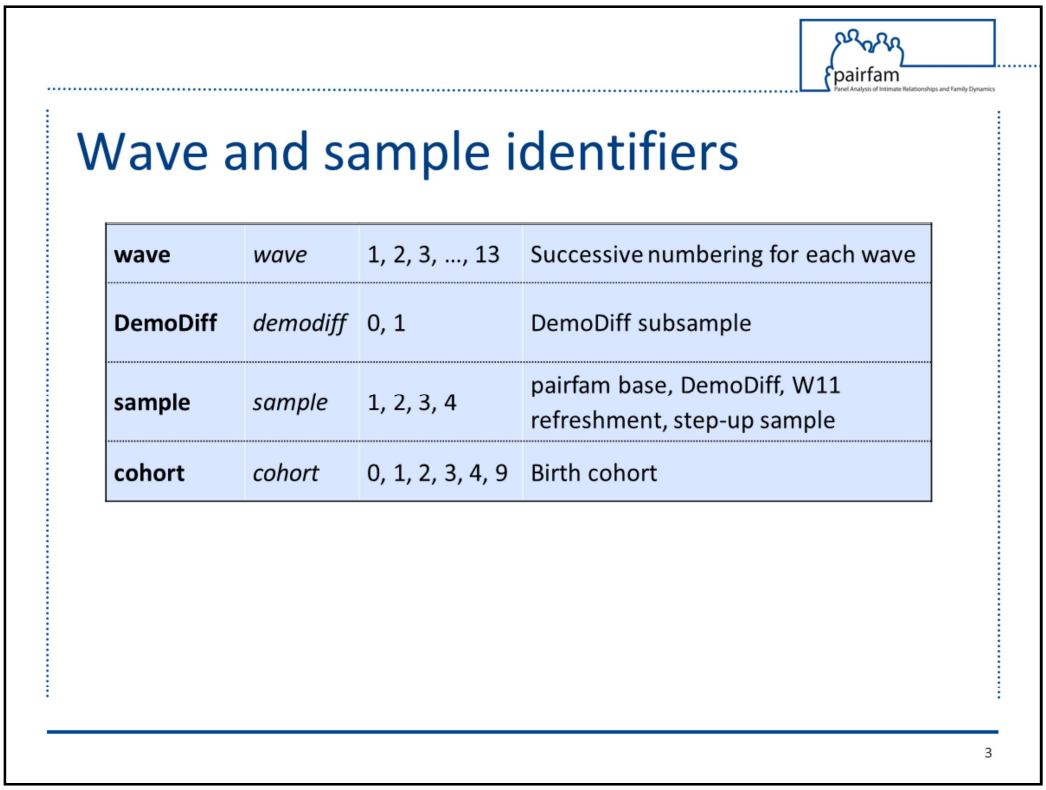
parent (the reference is always the anchor respondent). In case a child is an anchor's step child, *parentidstepk* contains the identification number of the child's biological mother or father. Note that these variables can differ from the variable *pid,* which denotes the anchor respondent's current partner. The other parent of the child might be a former partner.

For the anchor's parents, *mid* identifies the anchor respondent's mother (*id* + 301), *fid* the father (*id* + 302), *smid* for a stepmother (*id* + 303), and *sfid* for a stepfather (*id* + 304). If a new stepmother (or stepfather) was introduced, the next odd (or even) number was assigned.

The parent data include up to three parents per anchor respondent. Each parent provides information about their children – the anchor respondent and their siblings. Each sibling was also assigned a unique identifier, *sibid,* which is part of the parent data. "X" denotes the order of the siblings. Please be aware that the order of the anchor's siblings in the parent data is not identical to that in the anchor data.

Former child respondents who became anchor respondents themselves, so-called "step-up" respondents, receive a new *hhid* and *id* after completing the first interview as anchor respondents. The variable *f_cid* represents their former identifier as a respondent of the child interview (*cid*), making it possible to merge *step-up* data with their parents' anchor data or with their own former child data.

Further information on person identifiers can be found in the Data Manual.



How can you identify the different samples and waves?

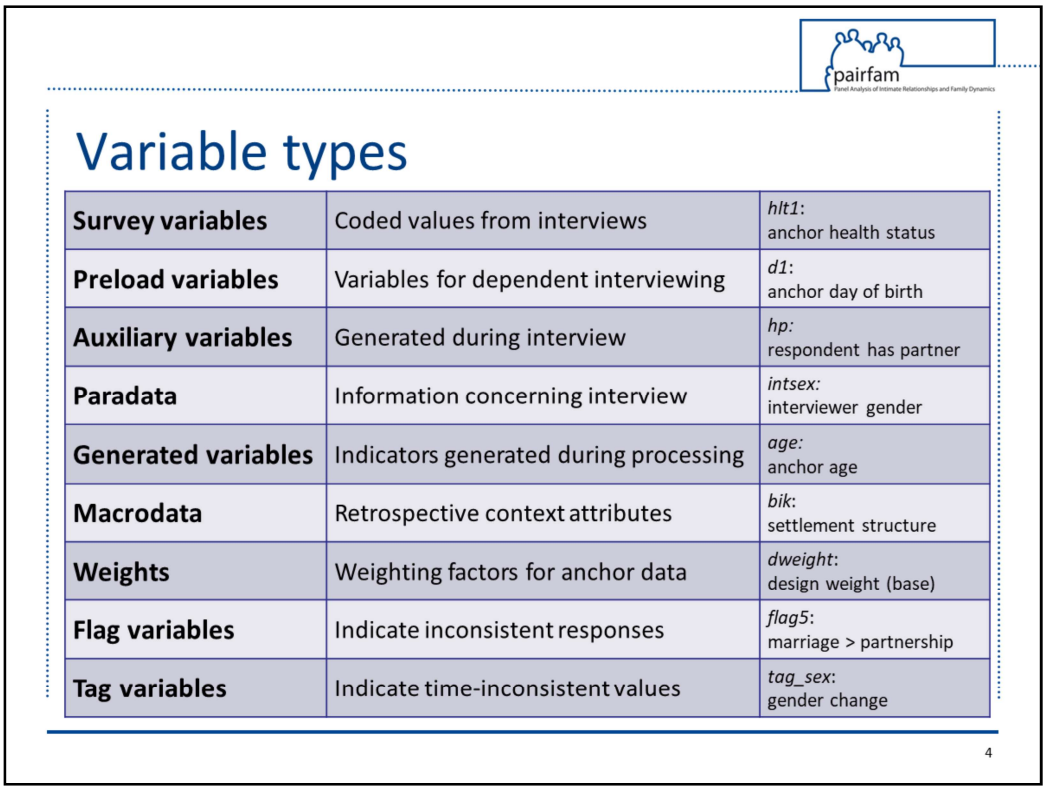
The variable *wave* is successively numbered for each fielding period, representing the survey wave.

The *pairfam* study consists of various anchor samples, which can be identified by three variables: *demodiff, sample, and cohort*.

Respondents from the integrated *DemoDiff* study can be identified by the *demodiff* variable.

The variable *sample* categorizes respondents from all samples: the *pairfam* base, *DemoDiff*, wave 11 refreshment, and *step-up* sample.

The generated variable *cohort* represents the birth cohort from which the anchor respondents were drawn. Values 1 to 4 represent the four birth cohorts from the base and refreshment sample. 0 and 9 represent *step-up* respondents, with 0 marking the former focus child's first interview and 9 for subsequent interviews.



Which types of variables are included in the anchor data?

Most of the data stored are survey variables, which are coded values provided by the respondents. For example, *hlt1* represents the anchor respondent's self-rated health status.

From wave 2 onward, the anchor data sets also contain preload variables. These are generated variables used for dependent interviewing. For example, *d1* stores the anchor's day of birth. This information is then pre-loaded in subsequent interviews to determine question routing and wording to improve the personalization of the questionnaire. Dependent interviewing is also used to ease the burden of the interview and to remind respondents of information reported in the last interview.

Auxiliary variables are generated during the interview and are also used to improve filtering and question wording. The variable *hp*, for example, stores whether the anchor respondent has a partner and determines which follow-up questions are posed to the anchor.

Paradata includes information concerning the interview situation and the interviewer. The variable *intsex*, for example, stores interviewer gender.

Generated variables are indicators created during data processing to facilitate user analyses. For example, the variable *age*: During the interview, the anchor respondent's day, month, and year of birth is collected. From this information, the variable *age* is then generated during data processing to store the anchor

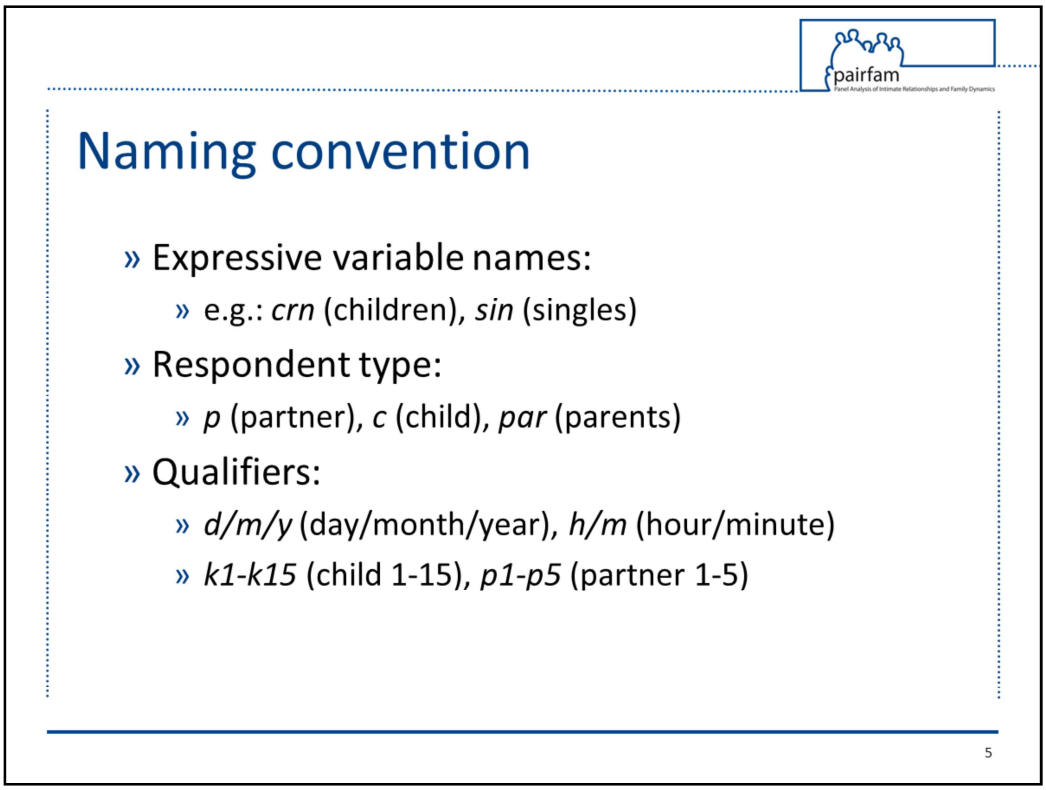
respondent's age at the time of each interview.

Macrodata are retrospective context attributes also generated during data processing. The variable *bik*, for example, provides information about the settlement structure of the anchor's main residence.

Weighting factors are also generated during data processing and are provided for the anchor data for each of the available samples: the *pairfam* base, refreshment sample, and integrated *DemoDiff* sample.

Flag variables indicate inconsistent responses within an interview. The variable *flag5*, for example, indicates that the anchor respondent reported being married to their partner before their relationship with the same partner began.

Tag variables indicate time-inconsistent values across survey waves. For example, *tag_sex* indicates a change in gender between waves.



Is there a variable naming convention?

The goal in naming the *pairfam* variables was to create expressive variable names, for example *crn* for the children module and *sin* for the module for singles.

Furthermore, responses from different respondent types are recognizable by a prefix – *p* for partner, *c* for child, and *par* for parents.

pairfam also implements qualifiers, for example *d* stands for day, *m* for month, and *y* for year, *h* for hour, and *m* for minutes.

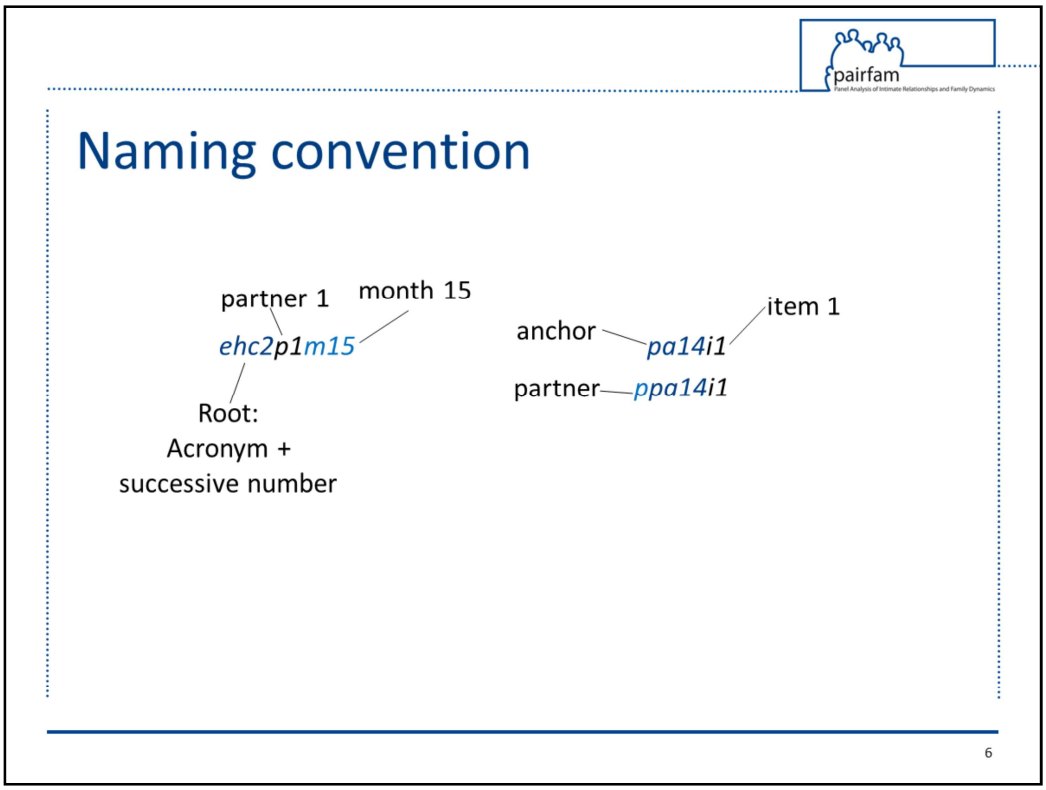
In addition, some anchor variables for responses to questions concerning children and partners include identifying information in the variable name. Children and partners are numbered consecutively as part of the variable name (_kx and_*px*).

All of the anchor's children are assigned a number the first time they are mentioned (*k1* to *k15*). The established numbering of the anchor's children remains constant throughout the survey, even in the case of death.

In contrast, the numbering of partners (*p1* to *p5*) as part of some variable names follows a rolling system and can be occupied by different partners in different waves. The rolling system works as follows: In wave 1, the current partner was assigned position *p0*. In wave 2, the pre-loaded partner information from wave 1 received the position *p1*, any new current partner was assigned to position *p2*, and partners reported between waves occupy the positions *p3* to *p5*.

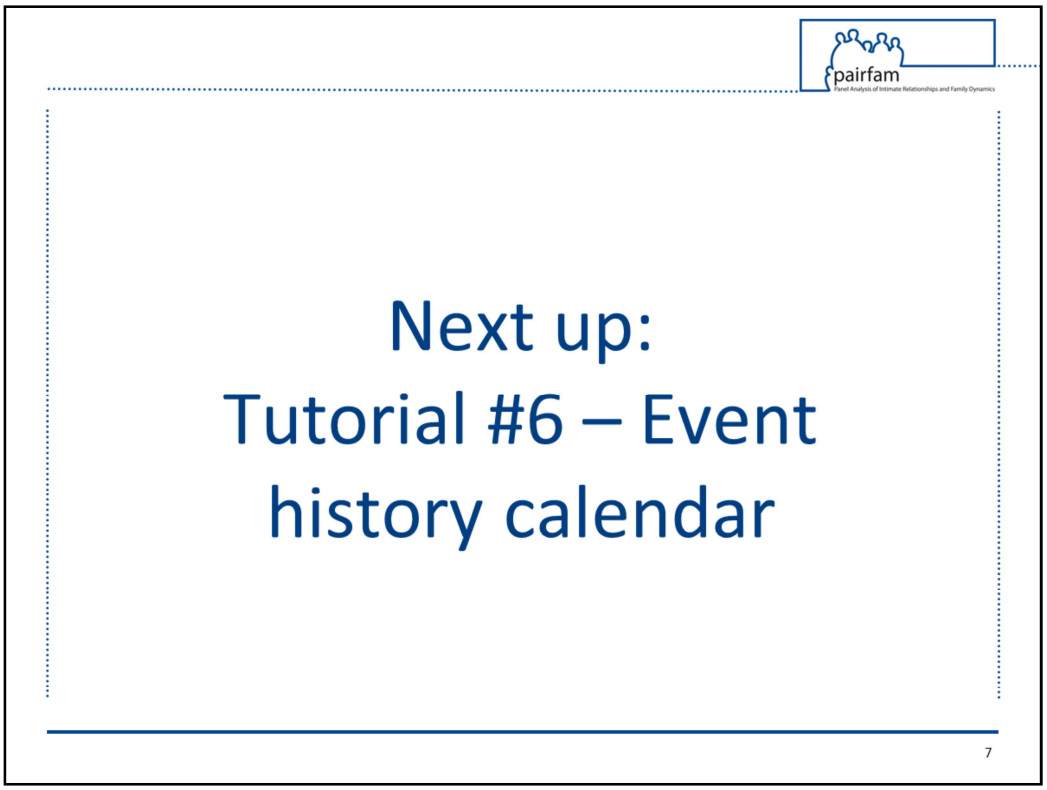
As of wave 3, the auxiliary variable *hpnr* in the anchor data sets contains the running

number of the current partner. Generally, information on partners is documented only for the current relationship up to one year after separation. The current partner can always be identified by the partner identifier *pid*, which remains stable throughout the panel.



Let's have a look at three examples. The variable *ehc2p1m15* stores the relationship status with partner 1 in month 15. The root of the variable is *ehc2*. It consists of an acronym that describes the topic of the variable (*ehc*) plus a number for the question in this module (*2*). *p1* stands for "partner 1" and *m15* stands for "month 15".

Now let's look at the variables *pa14i1* and *ppa14i1*. *pa14i1* includes the anchor's response to question 14 concerning housework division with their partner (*pa*). The anchor's partner received the same question in the partner survey, so to keep things simple the question was named the same and the prefix *p* was added for "partner". This convention makes it easy to merge anchor and alteri data sets without changing the name of the variables beforehand. The suffix *i1* indicates that this question belongs to an item battery and represents the first item.



This is the end of the fifth tutorial.

The next tutorial will explain the Event History Calendar, or “EHC” module, in more detail.