

---

Subject: What is the underlying criteria for merging datasets from various waves?

Posted by [sufi](#) on Thu, 09 Jun 2016 12:47:44 GMT

[View Forum Message](#) <> [Reply to Message](#)

---

Hello all,

I'm working on a large piece of data which contains virtually all African datasets in the DHS. I'm having some trouble figuring out the exact criteria for merging individual recode files with male recode, hiv, and GPS files.

I fully understand that matching individuals works at the cluster-householdno-personal number level. However, I'm unsure as to the criteria for merging files of the same country, in the same year.

Take Cameroon 2011 for example:

Would all files merge by cluster-householdno-personal number when using data from Cameroon 2011? The reason I'm unsure is because, for example, the hiv data is called CMAR61FL, while the male recode data is called CMMR60FL

One has the number '61' and the other has the number '60'. Should they still merge?

I would appreciate any help and clarification

Thanks in advance

---

---

Subject: Re: What is the underlying criteria for merging datasets from various waves?

Posted by [Bridgette-DHS](#) on Tue, 14 Jun 2016 14:28:18 GMT

[View Forum Message](#) <> [Reply to Message](#)

---

Following is a response from Senior DHS Stata Specialist, Tom Pullum:

The fifth character of the file name is the phase of DHS (0 through 7, but 0 refers to phase 1; we are currently in phase 7). The sixth character is the version. It is usually 0 for the first version of the data file, when the survey is first released, and then if corrections or additions (almost always of negligible importance) are made, the version is updated to 1, 2, etc. CMAR61FL and CMMR60FL are completely compatible. The AR file had one update but the MR file did not.

Sometimes this happens--one of the files from a survey has had more updates than another one of the files from the same survey. This should not affect a merge.

If there was a second survey in the same phase of DHS, then the first version of the second survey will be A, which is updated to B, etc. If there was a third survey, then the first version of the third survey will be G, which is updated to H, etc.

Obviously the file naming system is sometimes awkward. It goes back to a time when the file name (before the dot) could be at most 8 characters long. That's no longer a restriction, and I expect the whole system will be revised before long.....

---