Documentation for the Standardization of the Harmonized Histories Data File for USA for birth, partnership histories, leaving home questions and background variables

HARMONIZED HISTORIES USA NSFG 2007 (13495 respondents)

Karolin Kubisch
Max Planck Institute for Demographic Research Rostock

Katherine Michelmore
Cornell University

October 2010
Updated 2012
Updated 2014

The following documentation gives a description of all input variables and the consequent preparation of the output variables according to the manual for the preparation of comparative fertility and union histories.
All problem cases as well as the treatment of these cases are described in detail.

Missing values are coded:
.a unknown
.b does not apply
.c unavailable in survey

Source: NSFG 2007

Interview dates NSFG 2007: From January to December 2006 to 2008

Update 2012:

This file makes a few corrections to the previous file in the following ways: The primary problem with the previous file is that there were too many unions reported without accompanying dates, primarily for the men. The UNINUM variable was double-counting current cohabiting partners for men such that any man currently living with a partner had one extra union in the UNINUM variable. This also affected the UNION_$ variables since these variables are created based on the number of unions reported in UNINUM. The UNION_Y$ and UNION_M$ variables are unchanged.

In addition to double-counting the cohabiting partners of men currently living with a partner (not wife), many men reported more cohabiting unions than they provided dates for. This was not just a matter of not knowing the start and end dates, but rather that the respondent was never asked the start and end dates of some unions. I therefore altered the UNINUM variable such that it only includes unions that the respondent provided information on the dates of those unions, regardless if that information was 'refused' or 'don't know'.

This change alters the value of UNINUM for many men, particularly those with several unions. It also alters the UNION_$ variables. In the previous version of this file, many UNION_$ variables were provided without accompanying UNION_Y$ and UNION_M$ variables. This change reduces the number of UNION_$ without information on the dates.

Finally, there was a coding error in the previous file that mistakenly coded YEARBIRP_$ as missing for many individuals. The change made to UNINUM and UNION_$ variables fixed the problem with YEARBIRP_$ and there is now far less missing data for this variable. Again, this is primarily a change for the men in the 2007 file.

June 2014: Corrections in the variables to leaving home histories of children (KID_L, KID_LY, KID_LM)

---

1. Part Basic Information

RESPID: ID number to be assigned at merging LEAVE BLANK

ARID: ID number from raw data (original ID number) used: caseid
13495 respondents

COUNTRY: Country and survey
COUNTRY: code: 8402: USA NSFG 2007
no missing cases

MONTH_S: Month of survey used: cmintwv
codes: 1-12
no missing cases

IMONTH_S: Month of survey, including imputed dates
For missing values imputation:
randomly variable between 10 and 12

YEAR_S: Year of survey used: cmintwv
YEAR_S: 2006-2008
no missing cases

SEX: Sex of the respondent used: inmale, infem
No missing cases
Sex structure of the respondents:
Female: 7356
Male: 6139

BORN_Y: Year of birth of respondent used: cmbirth
1961-1993: no missing cases

BORN_M: Month of birth of respondent used: cmbirth
no missing cases

IBORN_M: Month of birth of respondent including imputed months
Randomly, variable between 1-12

---

2. Part LEAVING HOME

LEAVE_1: Indicator of whether "left home"

not available in survey

LEAVE_Y1: Year of first time leaving home

not available in survey

LEAVE_M1: Month of first time leaving home

not available in survey

ILEAVE_M1: Month of first time leaving home and imputed months:

not available in survey

---

3. Part UNIONS AND DISSOLUTION ($=order of union)

Marriages/Unions:

There were a couple of issues with the marriage/union histories. The NSFG computes variables such as marriage start and end dates, but the end dates include separation and divorce dates. So for couples that separate before they divorce, I only know the separation date, not the divorce date. There are raw variables that code for the actual divorce dates, but some of the values are spurious--divorce dates occur before the actual marriage occurs, so I decided to use the cleaned variable. I have indicated which dates have been imputed with the imputation variables described in the manual (e.g. IDIV_M). Individuals with imputed dates for marital dissolution do not have values for DIV_M, but they are included in the IDIV_M variable, which contains information on divorce dates regardless of whether the answer was imputed.

Calculation:
The biggest challenge for creating the union start and end dates was synthesizing the separate information about cohabiting partners and marital partners. The NSFG has three sets of variables for start and end dates of unions for respondents. There are variables for previous

cohabiting partners who they did not marry, marriage start and end dates for their first five marriages, and pre-marital cohabitation start dates for those marriages. I calculate the marriage and union histories by taking the earliest start date mentioned in any of the three sets of variables-- cohabitation start date, marriage start date, and per-marital cohabitation start date. Once I find the earliest date, I map on the rest of the information about the partner- the birth dates and the number of children they have had with previous mates. Once I gather all of the information from the first union/marriage, I set the raw start and end date variables equal to missing so they are no longer considered in future unions. I continue in this manner until all of the cohabitations and marriages have been enumerated. In 1995 the NSFG asks about 5 marriages and 7 cohabiting partners, and the 2007 NSFG asks about 5 marriages and 5 cohabiting partners for women, and 4 marriages and 3 previous partners for men.

Finally, the NSFG does ask all respondents how many times they have been married and how many cohabiting partners they have ever had, which is how I ascertained the total number of unions, UNINUM. It's possible, however, that the NSFG would not have collected information about all of these unions, so that the value for UNINUM may exceed the number of unions for which we have information on start and end dates, and birth dates. This is especially true for the 2007 males, who are not asked about any of their previous cohabitation partners aside from their first partner and any partners in the last 12 months. I thought it best to utilize all of the data we have though, so I let the UNINUM value include all previous unions, not just those with explicit information on start and end dates in the NSFG.

UNINUM: Total number of unions used: timesmar, hmothmen

Syntax:
Egen UNINUM=rsum(timesmar, hmothmen)

UNINUM (old):	UNINUM (March 2012)
0: 5008	5055
1: 5198	5717
2: 2035	2001
3: 746	536
4: 277	136
5: 93	38
6: 47	11
7: 28	1
8: 20	0
9: 43	0

In the old original data there were also unions from number 10 until 91.
The total number of unions was generated with:

```
gen UNINUM1=0
forvalues x=1/9 {
replace UNINUM1=UNINUM1+1 if UNION_`x'>0
            }
```

UNION_$: UNION order used: UNINUM

Definition (UNION_1 to UNION_x)
➔A union exists if the respondent reports at least x unions in the UNINUM variable.

UNION_1: 8440
UNION_2: 2723
UNION_3: 722
UNION_4: 186
UNION_5: 50
UNION_6: 12
UNION_7: 1
UNION_8: 0
UNION_9: 0

No missing cases

UNION_Y$: Year of start union used: cmcohstxx, mardat0x, cmpmcohxx

Filter: UNION_Yx=.b if UNION_x==0

No missing cases

UNION_M$: Month of start UNION

Filter: UNION_Mx=.b if UNION_x==0

No missing values

IUNION_M$: Month of start UNION
and imputed months

Filter: IUNION_Mx=.b if UNION_x==0

SEP_$: Dissolution of UNION used: cmstpcohxx, mardis0x

Filter: SEP_x=.b if UNION_x==0

No missing cases

Order of Union	Number of unions	number of separations	death of partner
1	UNION_1: 8440	4481	
2	UNION_2: 2723	1408	

3	UNION_3: 722	386	
4	UNION_4: 186	103	
5	UNION_5: 50	26	
6	UNION_6: 12	7	
7	UNION_7: 1	0	

SEP_Y$: Year of end of UNION

Filter: SEP_Yx=.b if UNION_x==0
SEP_Yx=.b if SEP_x==0

No missing cases

SEP_M$: Month of end of UNION

Filter: SEP_Mx=.b if UNION_x==0
SEP_Mx=.b if SEP_x==0

No missing cases

ISEP_M$: Month of end of UNION
and imputed months
according to manual page 4 (random)

Filter: ISEP_Mx=.b if UNION_x==0
ISEP_Mx=.b if SEP_x==0

---

4. Part MARRIAGE AND DIVORCE ($=order of union)

MARR_$: Indicator of whether marriage took place
and type of marriage

used: mardat0x

Filter: MARR_x=.b if UNION_x==0

No missing cases

Order of Union	Number of unions	number of marriages
1	UNION_1: 8440	4883
2	UNION_2: 2723	1337
3	UNION_3: 722	290
4	UNION_4: 186	69
5	UNION_5: 50	20
6	UNION_6: 12	2
7	UNION_7: 1	0

MARR_Y$: Year of marriage used: mardat0x

Filter: MARR_Yx=.b if UNION_x==0
MARR_Yx=.b if MARR_x==0

MARR_Y2 missing values: 10

MARR_Y3 missing values: 8
MARR_Y4 missing values: 3
MARR_Y5 missing values: 1

MARR_M$: Month of marriage used: mardat0x

Filter: MARR_Mx=.b if UNION_x==0
MARR_Mx=.b if MARR_x==0

MARR_Y2 missing values: 10
MARR_Y3 missing values: 8
MARR_Y4 missing values: 3
MARR_Y5 missing values: 1

IMARR_M$: Month of marriage
and imputed months
according to manual page 4 (random)

Filter: IMARR_Mx=.b if UNION_x==0
IMARR_Mx=.b if MARR_x==0

DIV_$: Indicator of whether divorce occurred used: marend0x

Filter: DIV_x=.b if UNION_x==0
DIV_x=.b if MARR_x==0

No missing cases

Order of Union	Number of unions	number of marriages	number of divorces
1	UNION_1: 8440	4883	1356
2	UNION_2: 2723	1337	364
3	UNION_3: 722	290	69
4	UNION_4: 186	69	16
5	UNION_5: 50	20	5
6	UNION_6: 12	2	1
7	UNION_7: 1	0	0

DIV_Y$: Year of divorce used: mardis0x

Filter: DIV_Yx=.b if UNION_x==0
DIV_Yx=.b if MARR_x==0
DIV_Yx=.b if DIV_X==0 or .d

No missing cases

DIV_M$: Month of divorce used: mardis0x

Filter: DIV_Mx=.b if UNION_x==0
DIV_Mx=.b if MARR_x==0
DIV_Mx=.b if DIV_x==0 or .d

DIV_M1 missing values: 27
DIV_M2 missing values: 10
DIV_M3 missing values: 2

IDIV_M$: Month of divorce
and imputed months
according to manual page 4 (random)

Filter: IDIV_Mx=.b if UNION_x==0
IDIV_Mx=.b if MARR_x==0
IDIV_Mx=.b if DIV_x==0 or .d

---

5. Part PARTNER`S CHARACTERISTICS ($=order of union)

SEXP_$: Partner`s sex used: infem,inmale

Filter: SEXP_x=.b if UNION_x==0

No missing cases

YEARBIRP_$: Year of birth of partner used: cmhsbdobx

Filter: YEARBIRP_x=.b if UNION_x==0

YEARBIRP_1 missing cases: 453
YEARBIRP_2 missing cases: 164
YEARBIRP_3 missing cases: 59
YEARBIRP_4 missing cases: 18
YEARBIRP_5 missing cases: 4
YEARBIRP_6 missing cases: 4
YEARBIRP_7 missing cases: 0

MONBIRP_$: Month of birth of partner used: cmhsbdobx

Filter: MONBIRP_x=.b if UNION_x==0

MONBIRP_1 missing cases: 453
MONBIRP_2 missing cases: 164
MONBIRP_3 missing cases: 61
MONBIRP_4 missing cases: 19
MONBIRP_5 missing cases: 4
MONBIRP_6 missing cases: 4
MONBIRP_7 missing cases: 0

IMONBIRP_$: Month of birth of partner
and imputed months
according to manual page 4 (random)

Filter: IMONBIRP_x=.b if UNION_x==0

NUMCHP_$: Number of children of partner used: numkdshx
at start of union$

NUMCHP_$ can only get at current partner`s kids.

Filter: NUMCHP_$=.b if UNION_X==0

NUMCHP_1: missing values: 2114
NUMCHP_2: missing values: 770
NUMCHP_3: missing values: 171
NUMCHP_4: missing values: 39
NUMCHP_5: missing values: 10
NUMCHP_6: missing values: 0
NUMCHP_7: missing values: 0

NUMCLIV_$: Number of children of partner lived with respondent

NUMCLIV_$ can only get at current partner`s kids.

No values

---

6. Part Birth histories (biological kids)

Birth histories:
For the birth histories, there is a separate file in the NSFG for each pregnancy a respondent reports.
This required reshaping the data to a respondent-level file, and enumerating each birth from there. I dropped all of the pregnancies that did not result in a live birth, and then collected the information necessary from there. I looped through up to 11 births to code for birth date, sex, leaving dates, and death dates. Multiple births are dealt with by copying the information from the one pregnancy onto a new line of data before the file is reshaped. I then copy the baby characteristics from baby number 2 into the baby number 1 spot so the code will work for all births. See the file for details.

KID_$: Indicator of child order used: biodob

no missing cases

Child order	number of children
1	6355
2	4103
3	1847
4	691
5	246
6	93
7	32
8	12
9	6
10	4
11	1

KID_Y$: Year of birth of child used: biodob

Filter: KID_Yx=.b if KID_x==0

KID_Y1 missing values: 16
KID_Y2 missing values: 18
KID_Y3 missing values: 14
KID_Y4 missing values: 18
KID_Y5 missing values: 12
KID_Y6 missing values: 6
KID_Y7 missing values: 4
KID_Y8 missing values: 2
KID_Y9 missing values: 2

KID_M$: Month of birth of child used: biodob

Filter: KID_Mx=.b if KID_x==0

KID_M1 missing values: 16
KID_M2 missing values: 18
KID_M3 missing values: 14
KID_M4 missing values: 18
KID_M5 missing values: 12
KID_M6 missing values: 6
KID_M7 missing values: 4
KID_M8 missing values: 2
KID_M9 missing values: 2

IKID_M$: Month of birth of child
and imputed months
according to manual page 4 (random)

Filter: IKID_M_x=.b if KID_x==0

KID_S$: Sex of child used: biosex

Filter: KID_Sx=.b if KID_x==0

KID_S1 missing values: 2
KID_S2 missing values: 3
KID_S3 missing values: 1
KID_S4 missing values: 3
KID_S5 missing values: 3
KID_S6 missing values: 2
KID_S7 missing values: 1

Child order	number of children	male	female
1	6355	3258	3095
2	4103	2067	2033
3	1847	930	916
4	691	351	337
5	246	126	117
6	93	51	40

7	32	18	13
8	12	6	6
9	6	2	4
10	4		4
11	1	1	

Death and leaving variables not ascertained in male file

KID_D$: Death of child used: ALIVENOWA

Filter: KID_Dx=.b if KID_x==0

missing cases:
KID_D1: 1
KID_D2: 1
KID_D3: 1
KID_D4: 1
KID_D5: 1
KID_D6: 1
KID_D7: 1

Child order	number of children	death
1	6355	35
2	4103	31
3	1847	15
4	691	5
5	246	2
6	93	3
7	32	
8	12	
9	6	
10	4	
11	1	

KID_DY$: Year of death of child used: CMKIDIEDA

Filter: KID_DYx=.b if KID_x==0
KID_DYx=.b if KID_Dx==0

No missing values

KID_DM$: Month of death of child used: CMKIDIEDA

Filter: KID_DMx=.b if KID_x==0
KID_DMx=.b if KID_Dx==0

No missing cases

IKID_DM$: Month of death of child
and imputed months

according to manual page 4 (random)

Filter: IKID_DMx=.b if KID_x==0
IKID_DMx=.b if KID_Dx==0

KID_L$: Child left home used: CMKIDLFTA

Filter: KID_Lx=.b if KID_x==0
2014: children which died were excluded from KID_L=1 and are now coded with special missing code .d and KID_LY and KID_LM for dead children is coded as .b.

KID_L1 missing cases: 7
KID_L2 missing cases: 3
KID_L3 missing cases: 3
KID_L4 missing cases: 3
KID_L5 missing cases: 4
KID_L6 missing cases: 3
KID_L7 missing cases: 1

Child order	number of children	Left home
1	6355	1101
2	4103	599
3	1847	296
4	691	129
5	246	52
6	93	28
7	32	11
8	12	4
9	6	3
10	4	2
11	1	0

KID_LY$: Year child left home used: CMKIDLFTA

Filter: KID_LYx=.b if KID_x==0
KID_LYx=.b if KID_Lx==0

KID_LY1 missing cases: 7
KID_LY2 missing cases: 3
KID_LY3 missing cases: 3
KID_LY4 missing cases: 3
KID_LY5 missing cases: 4
KID_LY6 missing cases: 3
KID_LY7 missing cases: 1

KID_LM$: Month child left home used: CMKIDLFTA

Filter: KID_LMx=.b if KID_x==0
KID_LMx=.b if KID_Lx==0

KID_LM1 missing cases: 7
KID_LM2 missing cases: 3
KID_LM3 missing cases: 3
KID_LM4 missing cases: 3
KID_LM5 missing cases: 4
KID_LM6 missing cases: 3
KID_LM7 missing cases: 1

IKID_LM$: Month of death of child

and imputed months

according to manual page 4 (random variable)

Filter: IKID_LMx=.b if KID_x==0
IKID_LMx=.b if KID_Lx==0

---

7. Part Education

Education:
Education histories were created using the highest degree attained variables in the NSFG. I classified individuals into the highest ISCED category (6) if the individual reported having a graduate degree, they received a 5 if they obtained a Bachelors' degree, a 4 if some college, but no degree, 3 if high school, 2 for lower-secondary, and 1 for just primary school. I then collapsed these categories by including all post-high school individuals in the top category, high school graduates in the middle, and less than high school in the bottom category. These can also be switched if a better model is proposed. I used a similar method for classifying the education of the respondent's parents. For the education degree dates, I only included dates for individuals who are no longer in school. To me, that made the most sense, and I recoded the EDU_Y to .b, not applicable. I can easily switch this though, if an actual date is preferred. On the 1995 file, there are completion years even if the respondent is still in school and in the 2007 file, we only know the completion dates for high school and Bachelors' degree.

There were some cases in the 2007 file of missing values for the year of completion dates-- respondents refused or didn't remember when they completed their highest level of education. All of these individuals had a high school degree or less, and I knew the actual grade completed. I used the following method for assigning completion dates for these individuals:
if the individual reported completing the 12th grade, I assigned their education completion year (IEDU_Y) to their birth year plus 18 years, the average age at completing 12th grade. If the respondent reported completing 11th grade, I added 17 years to their birth year for the education completion year. I continued in this way down to the 9th grade, which was the lowest grade reported of the individuals with missing education years. I assigned all

of the education months (IEDU_M) to 6, for June, as most schooling is finished in June in the United States.

INSCHOOL: Currently studying at the time of interview used: goschol

Currently studying: 3923 respondents
No missing cases

EDU_COU: Highest level of education, country specific used: hieduc

No missing cases

Definition:

The country specific codes include:
* a 3-digit country prefix(840)
* a 1-digit survey code (NSFG 2007=2) and
* a 2-digit country specific code for level of education (code 5-15 levels of education)

ISCED_7: Highest level of education
Achieved according to ISCED 1997 used: hieduc

Definition: ISCED_7=1 HIEDUC<=3
ISCED_7=2 HIEDUC >3 & <=8
ISCED_7=3 HIEDUC ==9
ISCED_7=4 HIEDUC==10,11
ISCED_7=5 HIEDUC==12
ISCED_7=6 HIEDUCY>=13

Harmonized:

ISCED	Number
0+1	
2	3867
3	3454
4	3508
5	1931
6	735

EDU_3: Highest level of education ISCED used: ISCED_7
Collapsed into 3 categories

Definition: High: ISCED_7=code 4,5,6
Medium: ISCED_7=code 3
Low: ISCED_7=code 1 or code 2

Level	Number
High	6174
medium	3454
low	3867
missing cases	

EDU_Y: Year highest level of education achieved used: cmbagrad

Missing cases: 4522

EDU_M: Month highest level of education achieved used: cmbagrad

Missing cases: 4522

IEDU_Y: Year highest level education achieved and imputed year

Definition for imputation: Imputation for year of education is described above. If the completion grade is known, the completion year is defined as the birth year plus the average age at which an individuals completes that grade.

Missing cases: 4164

IEDU_M: Month highest education achieved and imputed month

---

8. Part Background variables (ethnicity, nationality etc.)

NATIVE: Born in country used: brnout

Born in country: 11116
Born elsewhere: 2371
8 missing cases

ETHNOS: Ethnicity/nationality used: hisprace

Country specific variable (840+2+code)

No missing cases

BIRTH_COU: Country of birth

Country specific variable (840+2+code)

not available in survey

MIG_Y: Year of migration used: yrstrus

missing cases: 26

MIG_M: Month of migration

not available in survey

IMIG_M: Month of migration and imputed months

according to manual page 4 (random)

not available in survey

---

9. Part Background variables (parental background)

SIS_NO: Number of sisters

not available in survey

BRO_NO: Number of brothers

not available in survey

SIBS: Total number of sibs

not available in survey

SIS_DIED: Number of sisters that died

not available in survey

BRO_DIED: Number of brothers that died

not available in survey

ISCED_MO: Mother`s highest level of education used: momdegre

ISCED	Number
0+1	
2	3148
3	4416
4	2985
5	2634
6	
.b	106
missing	206

ISCED_FA: Father`s highest level of education used: daddegre

ISCED	Number
0+1	
2	2899
3	3967
4	2219
5	3034
6	
.b	1067
missing	309

EDU3_MO: Highest level of education of mother
ISCED 1997, collapsed into 3 categories used: ISCED_MO

Definition: 1 (high) if ISCED_MO=4,5,6
2 (medium) if ISCED_MO=3
3 (low) if ISCED_MO=1 or 2

Level	Number
High	5916
medium	4416
low	3148
.b	
missing cases	312

EDU3_FA: Highest level of education of father
ISCED 1997, collapsed into 3 categories used: ISCED_FA

Definition: 1 (high) if ISCED_FA=4,5,6
2 (medium) if ISCED_FA=3
3 (low) if ISCED_FA=1 or 2

Level	Number
High	5253
medium	3967
low	2899
.b	
missing cases	1376

WORK_MO: Mother`s occupation, when respondent was 15
Country codes

not available in survey

WORK_FA: Father`s occupation, when respondent was 15
Country codes

not available in survey

ISCO3_MO: Mother`s occupation, when respondent was 15
3 categories

not available in survey

ISCO3_FA: Father`s occupation, when respondent was 15
3 categories

not available in survey

NATIVE_MO: Mother born in country

not available in survey

NATIVE_FA: Father born in country

not available in survey

BIRTHCO_MO: Mother`s country of origin

not available in survey

BIRTHCO_FA: Father`s country of origin

not available in survey

PARDIVEV: Parents ever divorced/separated

not available in survey

PARDIV_15: Parents divorced before age of 15

not available in survey

---

10. Part Background variables (region, size of location)

REGION: Country region at time of interview

Country specific variable (840 +2 +code)

not available in survey

SIZE: Size of place of residence at time of interview,

Country specific variable (840+2+code)

not available in survey

ISIZE: Size of place of residence at time of interview

Standardized code

SIZE_15: Size of place of residence at age 15

not available in survey

ISIZE_15: Size of place of residence at age 15

Standardized code

---

11. Part Other background variables

RELIGION: Religious affiliation at time of interview

Country specific variable (840+2+code) used: religion

No missing cases

IRELIGION: Religious affiliation at time of interview

Standardized code

ADOPT: Number of adopted children of respondent used: adptotkd

FOSTER: Number of foster children of respondent used: othkdfos

STEP: Number of stepchildren of respondent used: relothkd

Number of children	Adopt	Foster	Step
1	55	43	160
2	8	9	51
3	1	1	11
4	2		4
5	1	3	1
6		1	1
7		1	
8			
9			
10		1	

12. Part Weights

HHWGT: Household weight - not available in survey

PERSWGT: Personal weight - used: finalwgt30

KISHWGT: Kishweight - not available in survey