Cross-Cohort Harmonization Dataset

This is the beta release of a dataset to harmonize NLS data across the various cohorts. **For now, we have harmonized data available for the NLSY79 and NLSY97.** We see three benefits in this harmonization project:

- It will make it easier to perform cross-cohort comparisons.
- Researchers doing more complex analyses can use data from this system as background independent variables with minimal added work. This will allow them to concentrate on more complicated variables for their analysis. It will insure that variables across cohorts are calculated in as close to the same fashion as possible.
- By providing standardized variables across cohorts, we hope to provide an entry for novice NLS users. We hope this will get more NLS data into classrooms.

The idea is that in the longer run we may have a set of variables that are comparable across all the cohorts and could be simply selected into one dataset for analysis with a minimum of effort. Included would be documentation that would explain differences across surveys as well as across rounds within the same survey.

Methodology

This dataset contains a subset of all variables in the datasets: at present only around 15, although in the future there could be over 100. Many are created variables. The goal is to have variables that can be used directly or with minimal manipulation in an analysis. **All tell status as of the interview date.** Analyses of data on a different date (say on the 29th birthday or the date of first marriage) is beyond the scope of this dataset and would have to be calculated by the researcher.

ID numbers have been standardized to 7 digits by adding a two-digit prefix that represents the survey. For example, ID 9534 in the NLSY79 is 7909534 in this dataset. This makes merging in data from the individual survey datasets fairly easy.

Variables that are not one-time variables are indexed by year and age. Since it is possible for sample members to be interviewed twice in the same year or at the same age, **year and age are standardized. Year is the year in which interviewing started for that round of data collection. Age is year in which interviewing started minus (year of birth + 1), e.g., 1997 – (1984 + 1) = 12.** Tables 1 (NLSY97) and 2 (NLSY79) display how this assumption works. For example, all NLSY97 respondents born in 1980 will have their round 1 interview data under age 16 and year 1997, even though they may have turned 17 by the interview date, or have been interviewed in 1998. This age and year breakdown facilitates comparative work across cohorts in a particular year or at comparable ages. Variable names are standardized across cohorts with the ending indicating from which year or which age the observation comes. For example, MARST99 is marital status in 1999, while MARSTA24 is marital status at age 24.

Table 1. Assumed Age and Year in Harmonization: NLSY97

Interview round and year	Year of Birth				
	1980	1981	1982	1983	1984
Round 1, 1997	16	15	14	13	12
Round 2, 1998	17	16	15	14	13
Round 3, 1999	18	17	16	15	14
Round 4, 2000	19	18	17	16	15
Round 5, 2001	20	19	18	17	16
Round 6, 2002	21	20	19	18	17
Round 7, 2003	22	21	20	19	18
Round 8, 2004	23	22	21	20	19
Round 9, 2005	24	23	22	21	20
Round 10, 2006	25	24	23	22	21
Round 11, 2007	26	25	24	23	22
Round 12, 2008	27	26	25	24	23
Round 13, 2009	28	27	26	25	24
Round 14, 2010	29	28	27	26	25
Round 15, 2011	30	29	28	27	26
Round 16, 2013	32	31	30	29	28
Round 17, 2015	34	33	32	31	30

Table 2. Assumed Age and Year in Harmonization: NLSY79

Interview round and year	Year of Birth							
	1957	1958	1959	1960	1961	1962	1963	1964
Round 1, 1979	21	20	19	18	17	16	15	14
Round 2, 1980	22	21	20	19	18	17	16	15
Round 3, 1981	23	22	21	20	19	18	17	16
Round 4, 1982	24	23	22	21	20	19	18	17
Round 5, 1983	25	24	23	22	21	20	19	18
Round 6, 1984	26	25	24	23	22	21	20	19
Round 7, 1985	27	26	25	24	23	22	21	20
Round 8, 1986	28	27	26	25	24	23	22	21
Round 9, 1987	29	28	27	26	25	24	23	22
Round 10, 1988	30	29	28	27	26	25	24	23
Round 11, 1989	31	30	29	28	27	26	25	24
Round 12, 1990	32	31	30	29	28	27	26	25
Round 13, 1991	33	32	31	30	29	28	27	26
Round 14, 1992	34	33	32	31	30	29	28	27
Round 15, 1993	35	34	33	32	31	30	29	28
Round 16, 1994	36	35	34	33	32	31	30	29
Round 17, 1996	38	37	36	35	34	33	32	31
Round 18, 1998	40	39	38	37	36	35	34	33
Round 19, 2000	42	41	40	39	38	37	36	35

Round 20, 2002	44	43	42	41	40	39	38	37
Round 21, 2004	46	45	44	43	42	41	40	39
Round 22, 2006	48	47	46	45	44	43	42	41
Round 23, 2008	50	49	48	47	46	45	44	43
Round 24, 2010	52	51	50	49	48	47	46	45
Round 25, 2012	54	53	52	51	50	49	48	47
Round 26, 2014	56	55	54	53	52	51	50	49

Variables

Table 3 displays the background variables for the NLSY79 and NLSY97 harmonization beta release.

Table 3: Harmonization Variables, Fixed Background Variables

CASEID cross-cohort identification code
COHORT 79 or 97
SAMPLE_SEX from round 1 1 Male 2 Female
SAMPLE_RACE from round 1, NLSY79 has no mixed race 1 Black 2 Hispanic 3 Non-Black / Non-Hispanic 4 Mixed race (Non-Hispanic)
BIRTHDATE~M Birth Date – month
BIRTHDATE~Y Birth Date – year
AFQT Armed Forces Qualifying Test percentile score Variable AFQT_3 from the NLSY79 and ASVAB_MATH_VERBAL_SCORE_PCT from the NLSY97.

Table 4 shows the variables that vary by age and year for the NLSY79 and NLSY97 harmonization beta release.

Table 4: Harmonization Variables, by Age and Year

	Age	Year
INTDATE~M Interview month	X	X
INTDATE~Y Interview year	X	X
RFNI Reason for non-interview 0 Interviewed 1 Refusal 2 Not able to locate 3 Deceased 4 Not fielded due to prior refusals 5 In sample that was dropped	X	X

6 Other		
AGEMONTHS Age in months at interview date It is calculated by subtracting the month of birth from the month of the interview and adding 12 times the difference in year between birth and interview. The day of the month of birth or interview is not used to maintain confidentiality.	X	X
MARSTAT Marital Status at interview date 0 Never-married 1 Married 2 Separated 3 Divorced 4 Widowed	X	X
HIGRATT Highest Grade attended at interview date	X	X
HIGRCOMP Highest Grade completed at interview date These are from self-reports, and are not the edited created variables for highest grade completed also available in both datasets.	X	X
EMPSTAT Employment status at interview date 1 EMPLOYED 0 NO INFO REPORTED FOR WEEK 2 NOT WORKING (UNEMP V. OLF NOT DETERMINED) 3 ASSOC. WITH EMP, GAP DATES MISSING, ALL TIME NOT ACCTD FOR 4 UNEMPLOYED 5 OUT OF LABOR FORCE 7 ACTIVE MILITARY SERVICE EMP_STAT variables are created using the work history arrays which are loaded with week-by-week records of the respondent's labor force status.	X	X

***NOTE**: Employment status (EMPSTAT) for younger ages/early interview years. In the NLSY79, weekly employment history arrays mostly begin at age 16 (or January 1, 1978 if older than 16 in first round). In the NLSY97, weekly employment history arrays mostly begin at age 14. Therefore, although there may be information in the weekly employment history arrays for younger ages or years that are used to make the variable EMPSTAT, it should be used with caution as most respondents won't have that information for young ages/early years.*