MIDUS
Midlife in the United States
A National Study of Health & Well-Being

Conventions for Creating MIDUS Refresher 2 Datasets

The MIDUS Administrative Core has developed conventions for creating MIDUS datasets for the MIDUS Refresher 2. Such conventions are necessary for efficient and accurate data merges across Projects 1 through 5 and across different samples or waves of data (MIDUS 1 through 3, Refresher 1 and 2, Milwaukee 1 through 2). Further, the introduction of metadata standards such as the Data Documentation Initiative (DDI), which MIDUS adheres to, make the conventions an integral part of accurate documentation of the MIDUS study.

The attached pages provide specific guidelines for naming, labeling and formatting variables in the MIDUS Refresher 2. We have also included coding conventions for variables, missing value designations, and guidelines for date and time variable formats.

I. File Naming Conventions

File naming conventions help manage and organize MIDUS. These conventions become increasingly more useful as MIDUS becomes more complex. For all file types, MIDUS will use the prefixes MR2 to designate the Refresher 2, followed by an underscore and the project number. Also, end the file name with a date stamp, two examples of which follow:

Example: Documentations/Instruments
Refresher 2: MR2_P1_PhoneInstrument_20250801

Data files should include additional information on the number of cases (and avoid using special characters such as =, &, %, etc.):

Example: Datasets
Refresher 2: MR2_P1_SURVEY_N2154_20250801

II. Variable Naming Conventions

Rationale:

- Metadata best practices support a consistent and simple variable naming scheme. Not only does this reinforce the continuity of longitudinal data, but it makes cleaning and programming new variables more efficient and ensures compatibility across different software platforms.
- The original naming conventions were adopted in 2004 when there were strict character limits on variable names in statistical software. While current software programs are

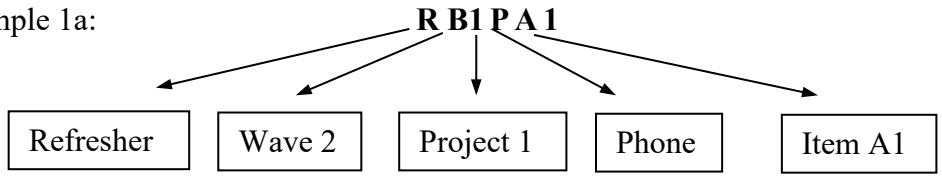
much more lenient in this regard, there are still substantial differences across statistical programs, and some older versions of statistical software still adhere to smaller variable name character limits.

- For these reasons, we will continue to limit variable name size, but because the Refresher cohort requires an "R" as the first character to identify the new sample (see examples 1a and 1b below), a 9-character variable name limit will be used.
- For the Refresher 2, the first 4 characters of each variable name will identify the cohort, longitudinal wave, the MIDUS project, and the instrument used to collect the data. The remaining characters identify the specific item or scale score variable that is represented by the measure's name. The exception to these conventions is the project 1 Milwaukee data. The Milwaukee sample is new at M2 and a used a different instrument (a personal interview instead of a phone interview) to collect Project 1 survey data from these individuals. Thus, Project 1 variables for the Milwaukee data include an additional character "A" to designate the project.[1]

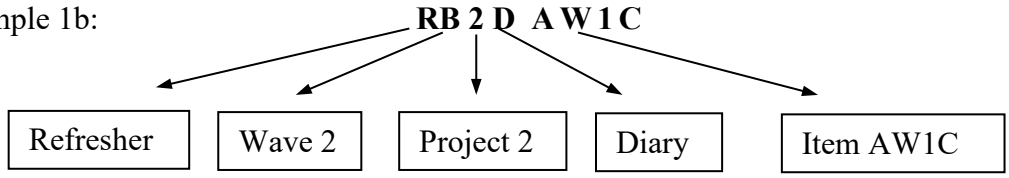
Examples: MIDUS Refresher 2

For the MIDUS Refresher 2, the second character of each variable name will be **B** for Wave 2. The extant naming conventions apply, i.e. those developed for M2.

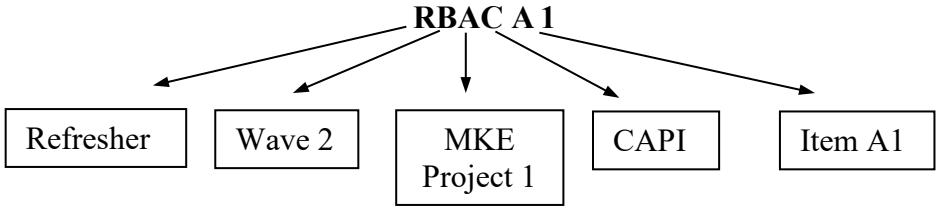
Example 1a:



Example 1b:



Example for Milwaukee Project 1:



[1] Moving forward, the MKE2 variable names continue the convention adopted at wave 2; MKE1 baseline was begun in 2005 during M2 data collection and so adopted the "B" character to indicate Wave2. Likewise, other projects like biomarkers and neuroscience who began baseline data collection at Wave2 also adopted variable names that began with "B". For MKE2, variable names mimic those of M3 with "C" as the first character.

III. Variable Labeling Conventions

Additional information about variables can (and should!) be included in the variable label. The label is the appropriate metadata field to more fully and clearly describe a variable. New technological metadata standards can use the rich information contained in a label to harvest, search, and identify specific variables. We are setting an 80-character/space limit for variable labels and encourage the use of mixed case text for more sentence-like descriptions of variables. See examples below.

Example 1a:

Variable name: **RB1PA4**
Variable label: **Days unable to work because of health (30 days)**

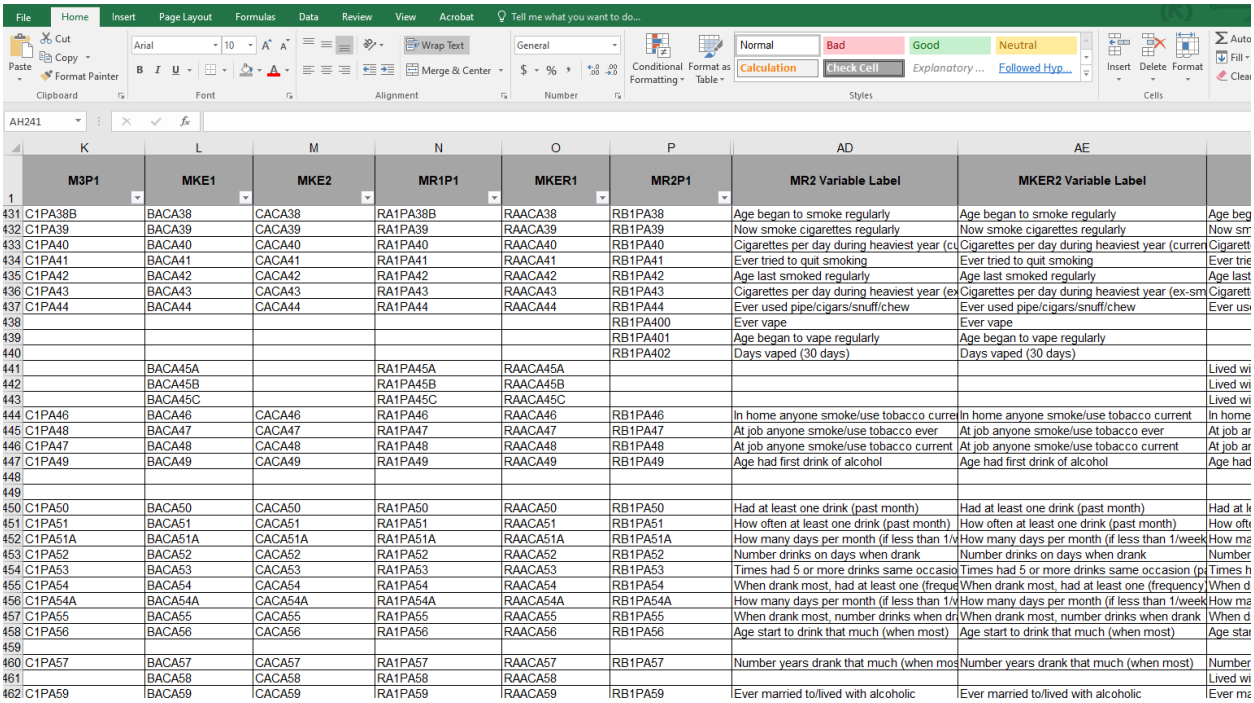
Example 1b:

Variable name: **RB4QCESD**
Variable label: **CESD: Center for Epidemiologic Studies Depression Scale**

IV. Variable concordance tables

The increasing number of waves and samples in MIDUS can make navigating among the datasets a challenge. We strongly suggest that each MIDUS project create variable concordance or cross-walk tables similar to the Excel spreadsheet created for Project 1 (see Figure 1 below). These tables not only help researchers find related variables, but also can be used by DDI codebook applications to facilitate variable searches across different MIDUS datasets. See the "Explore" and "Concordance Variables" views in the online MIDUS Portal (http://midus.colectica.org/Explore) to see how DDI applications make use of such concordance tables. Contact midus_help@aging.wisc.edu for more details if needed.

Figure 1: MIDUS Concordance Table



	K	L	M	N	O	P	AD	AE	
1	M3P1	MKE1	MKE2	MR1P1	MKER1	MR2P1	MR2 Variable Label	MKER2 Variable Label	
431	C1PA38B	BACA38	CACA38	RA1PA38B	RAACA38	RB1PA38	Age began to smoke regularly	Age began to smoke regularly	Age beg
432	C1PA39	BACA39	CACA39	RA1PA39	RAACA39	RB1PA39	Now smoke cigarettes regularly	Now smoke cigarettes regularly	Now sm
433	C1PA40	BACA40	CACA40	RA1PA40	RAACA40	RB1PA40	Cigarettes per day during heaviest year (cu	Cigarettes per day during heaviest year (current	Cigarett
434	C1PA41	BACA41	CACA41	RA1PA41	RAACA41	RB1PA41	Ever tried to quit smoking	Ever tried to quit smoking	Ever trie
435	C1PA42	BACA42	CACA42	RA1PA42	RAACA42	RB1PA42	Age last smoked regularly	Age last smoked regularly	Age last
436	C1PA43	BACA43	CACA43	RA1PA43	RAACA43	RB1PA43	Cigarettes per day during heaviest year (ex	Cigarettes per day during heaviest year (ex-sm	Cigarett
437	C1PA44	BACA44	CACA44	RA1PA44	RAACA44	RB1PA44	Ever used pipe/cigars/snuff/chew	Ever used pipe/cigars/snuff/chew	Ever us
438						RB1PA400	Ever vape	Ever vape	
439						RB1PA401	Age began to vape regularly	Age began to vape regularly	
440						RB1PA402	Days vaped (30 days)	Days vaped (30 days)	
441		BACA45A		RA1PA45A	RAACA45A				Lived wi
442		BACA45B		RA1PA45B	RAACA45B				Lived wi
443		BACA45C		RA1PA45C	RAACA45C				Lived wi
444	C1PA46	BACA46	CACA46	RA1PA46	RAACA46	RB1PA46	In home anyone smoke/use tobacco curre	In home anyone smoke/use tobacco current	In home
445	C1PA48	BACA47	CACA47	RA1PA47	RAACA47	RB1PA47	At job anyone smoke/use tobacco ever	At job anyone smoke/use tobacco ever	At job ar
446	C1PA47	BACA48	CACA48	RA1PA48	RAACA48	RB1PA48	At job anyone smoke/use tobacco current	At job anyone smoke/use tobacco current	At job ar
447	C1PA49	BACA49	CACA49	RA1PA49	RAACA49	RB1PA49	Age had first drink of alcohol	Age had first drink of alcohol	Age had
448									
449									
450	C1PA50	BACA50	CACA50	RA1PA50	RAACA50	RB1PA50	Had at least one drink (past month)	Had at least one drink (past month)	Had at l
451	C1PA51	BACA51	CACA51	RA1PA51	RAACA51	RB1PA51	How often at least one drink (past month)	How often at least one drink (past month)	How oft
452	C1PA51A	BACA51A	CACA51A	RA1PA51A	RAACA51A	RB1PA51A	How many days per month (if less than 1/w	How many days per month (if less than 1/week	How ma
453	C1PA52	BACA52	CACA52	RA1PA52	RAACA52	RB1PA52	Number drinks on days when drank	Number drinks on days when drank	Number
454	C1PA53	BACA53	CACA53	RA1PA53	RAACA53	RB1PA53	Times had 5 or more drinks same occasio	Times had 5 or more drinks same occasion (p	Times h
455	C1PA54	BACA54	CACA54	RA1PA54	RAACA54	RB1PA54	When drank most, had at least one (freque	When drank most, had at least one (frequency	When d
456	C1PA54A	BACA54A	CACA54A	RA1PA54A	RAACA54A	RB1PA54A	How many days per month (if less than 1/w	How many days per month (if less than 1/week	How ma
457	C1PA55	BACA55	CACA55	RA1PA55	RAACA55	RB1PA55	When drank most, number drinks when dr	When drank most, number drinks when drank	When d
458	C1PA56	BACA56	CACA56	RA1PA56	RAACA56	RB1PA56	Age start to drink that much (when most)	Age start to drink that much (when most)	Age sta
459									
460	C1PA57	BACA57	CACA57	RA1PA57	RAACA57	RB1PA57	Number years drank that much (when mos	Number years drank that much (when most)	Number
461		BACA58	CACA58	RA1PA58	RAACA58				Lived wi
462	C1PA59	BACA59	CACA59	RA1PA59	RAACA59	RB1PA59	Ever married to/lived with alcoholic	Ever married to/lived with alcoholic	Ever ma

V. Variable Formats

1. Utilize "Numeric" whenever possible; avoid using raw string variables. Open-ended responses, text, and verbatim data should be numerically coded where possible. Raw text or qualitative data can be formatted as a separate text file or spreadsheet.

2. Variable formats should be precise – variable lengths should not exceed the maximum number of digits possible for a response. Thus, if a response code has a maximum of 2 digits (e.g., a scale ranging from 1-10) then the variable length should be formatted as 2 digits.

3. Decimals – Specify up to 3 decimal places (an ICPSR convention). If important details of the data require more decimal places, please contact midus_help@aging.wisc.edu.

4. Date/time formats:

- Because of conflicting or proprietary formatting, date or time data provided **by *respondents*** must be separated into individual components. For example, date information must be recorded in separate month and year variables, and temporal information must be recorded in separate hour, minute and meridian (am/pm etc.) variables (one alternative for temporal variables is using a 24-hour clock or military time, in which hours and minutes can be represented as a numeric variable HHMM). Treating such variables this way allows them to be read by a wider array of software programs with fewer formatting problems or errors.
- Dates or times recorded **by *project staff for administrative purposes*** (e.g. date & time Medical History was completed) can be submitted to the Core in the date and time formats specified in the established conventions for MIDUS (e.g., mm/dd/yyyy, or hh:mm:ss, etc.).
- **Note: because of confidentiality/disclosure precautions, ICPSR suggests that month and year are sufficient for most sensitive date variables.**

5. Derived or constructed variables: if possible, any constructed or derived variables (e.g., scales scores or summary variables) are placed directly after their components in the dataset. That is, derived variables should follow their constituent variables in the sequence of variables in a dataset. Also, the details of the construction or derivation of such variables are to be explained in accompanying documentation (e.g., a Word document explaining the formula, procedure, source, criteria, etc., used in creating the constructed or derived variables).

VI. Value Labeling Conventions

A. Format

All value labels should be **UPPER CASE**.

Example:

1 = YES
2 = NO
7 = DON'T KNOW
8 = REFUSED/MISSING
9 = INAPP

B. Coding Conventions for Non-response (Don't Know, Missing Data, Inapplicable, Filters)

The following values should be used to indicate different types of non-response to questions or data fields (a series of 9's can be used as place holders):

DON'T KNOW (7's) – 7, 97, 997, 9997, etc.
Used to indicate explicit "Don't know" responses (where a specific response option of "Don't know" was offered).

REFUSED/MISSING (8's) – 8, 98, 998, 9998, etc.
Used to indicate R did not provide a response to a particular question.

INAPPLICABLE (9's) – 9, 99, 999, 9999, etc.
Used to indicate that R was not asked a particular question. This will occur most often because of skip patterns that the R is asked to follow, or questions that R determines do not apply to him or her.

INCOMPLETE SAQ (-1)
Used to indicate SAQ non-responders in the Project 1 protocol. The SAQ variables for those cases that did not return a completed SAQ will be coded -1.

C. Additional Coding Situations.

In M2, some projects used additional codes to indicate invalid values or incomplete data. We suggest using the value of 96 (and working backwards to add additional codes). For example, Project 2 assigned additional codes to cortisol variables that did not contain valid data (e.g., 96 = empty vial, 95 = not done, 94 = unreliable). Variable values should be blank or system-missing only if an R did not complete a whole instrument within a project's protocol. For example, in Project 1 individuals who do not complete the SAQs are assigned system-missing values for all SAQ variables. This is consistent with procedures followed in MIDUS 1 and 2.