

Conventions for Creating MIDUS Datasets

The MIDUS Administrative Core has developed conventions for creating MIDUS datasets. These conventions have evolved over time to accommodate technological advances and data management best practices. Such conventions ensure data quality, provide a similar look-and-feel to all MIDUS datasets, and facilitate efficient and accurate data merges across different MIDUS projects, samples, or waves. Further, the introduction of technological 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 MIDUS. Also included are coding conventions for variables, missing value designations, and guidelines for date and time variable formats.

Note: All project leaders will be responsible for delivering cleaned, coded *SPSS* data files to the Administrative Core. Accompanying text documentation must be in *Word* format¹ (ICPSR will create final PDFs for public release). We recommend sending an early draft of the dataset and documentation to the Administrative Core for review before you make final data deliveries.

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 MR and M3 to designate the Refresher and MIDUS 3, respectively, followed by an underscore and the project number. Avoid using special characters (" / \ : * ? " < > [] & \$) in filenames and all research metadata. These characters have specific meanings in computer operating systems and different applications that could cause problems. Similarly, use underscores (_) or type in CamelCase to separate terms, not spaces. MIDUS recommends including a date using the format recommended by International Standards Organization (ISO) 8601: YYYY-MM-DD. Finally, end the file name with a timestamp, two examples of which follow:

Examples: Documentation/Instruments

Refresher: MR_P1_PHONE INSTRUMENT_5-8-12

MIDUS 3: M3_P1_DocumentationOfScales_20120508

¹ When at all possible, include the date in the document footer or on the title page. This ensures that important versioning information is not removed by ICPSR upon submission. ICPSR renames submitted files and removes the MIDUS timestamp from the filename.

Data files should include additional information on the number of cases:

Examples: Datasets

Refresher: MR_P1_DATA_N2100_5-8-12

MIDUS 3: M3_P1_DATA_N5000_20120508

II. Variable Naming Conventions

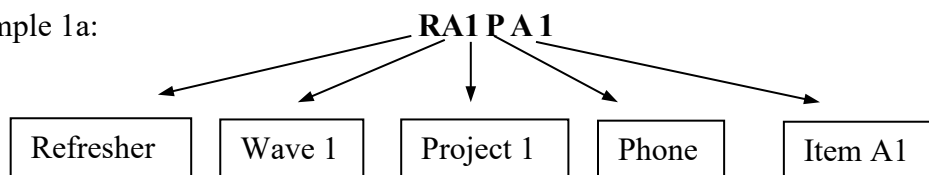
Rationale:

- 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 waves of 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 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 is used for MR variables.
- For the Refresher, the first 4 characters of each variable name identifies 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. MIDUS 3 follows the same conventions but uses the first 3 characters to identify wave, project, and instrument. 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.²

Examples: MIDUS Refresher

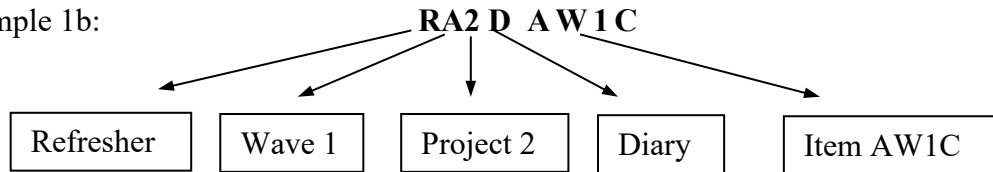
For the MIDUS Refresher the first character of each variables name will be **R**. Otherwise, the extant naming conventions apply, i.e. those developed for M2.

Example 1a:

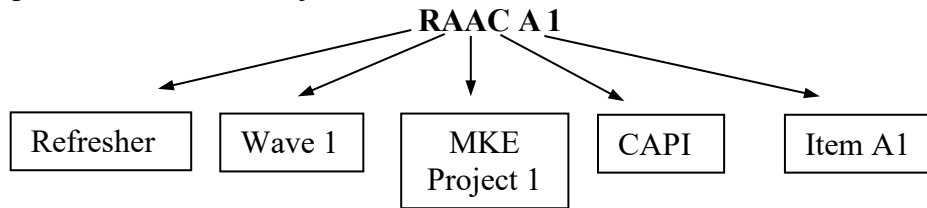


² Moving forward, the MKE2 variable names will continue the convention adopted at wave 2; MKE1 baseline was begun in 2005 during M2 data collection and so adopted the “B” character to indicate Wave2. Likewise, other projects like biomarkers and neuroscience who began baseline data collection at Wave2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

Example 1b:

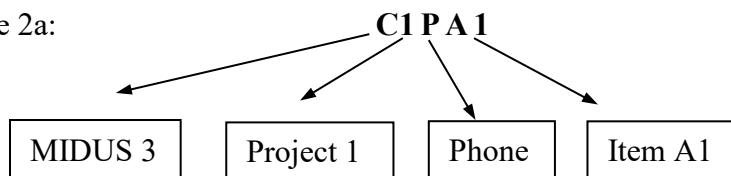


Example for Milwaukee Project 1:

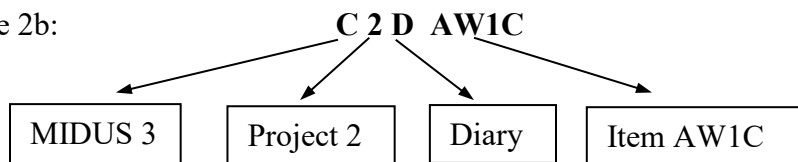


Examples: MIDUS 3

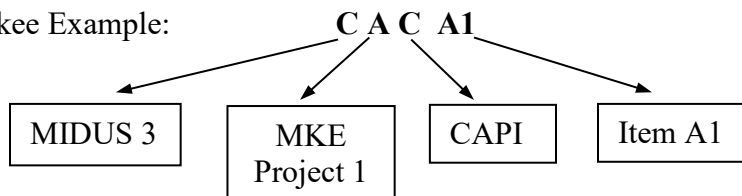
Example 2a:



Example 2b:



Milwaukee Example:



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: **RA1PA4**

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

Example 1b:

Variable name: **RA4QCESD**

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 by Project 1 (see example in Figure 1 below). These tables help researchers find related variables across datasets and can be used by Data Documentation Initiative (DDI) tools to facilitate online variable concordance and searches. See the “Explore” and “Concordance Variables” views in the online MIDUS Portal (<http://midus.colectica.org/Explore>) to see how DDI tools use such concordance tables. Contact Barry Radler (bradler@wisc.edu) for more details if needed.

Figure 1: MIDUS Concordance Table

| | A | B | C | D | E | F | G |
|----|------------------|------------------|-------------------|-------------------------|--|------------------------------------|---------------------------|
| | M2 Variable Name | M1 Variable Name | MKE Variable Name | Refresher Variable Name | M2 Variable Labels | M1 Variable Labels | Longitudinal Version Note |
| 25 | | | | | HEALTH | | |
| 26 | B1PA1 | A1PA4 | BACA1 | RA1PA1 | Physical health self-evaluated | Physical health | |
| 27 | B1PA2 | A1PA5 | BACA2 | RA1PA2 | Mental/emotional health self-evaluated | Mental or emotional health | |
| 28 | B1PA3 | A1PA6 | BACA3 | RA1PA3 | Health compared to others your age | Self-evaluated health | |
| 29 | B1PA4 | A1PA7 | BACA4 | RA1PA4 | Days unable to work b/c health (30 | Days work limited by health | |
| 30 | B1PA4A | A1PA7A | BACA4A | RA1PA4A | Reason unable to work (phys, ment | Unable to work, physical, mental | |
| 31 | B1PA4BA | A1PA7BA | BACA4BA | RA1PA4BA | Num days unable due to phys hlth c | # of Days physical | |
| 32 | B1PA4BB | A1PA7BB | BACA4BB | RA1PA2BB | Num days unable due to ment hlth c | # of Days mental | |
| 33 | B1PA4BC | A1PA7BC | BACA4BC | RA1PA2BC | Num days unable work due to ment | # of Days combination | |
| 34 | B1PA5 | A1PA8 | BACA5 | B1PA5 | Days cut back work b/c health (30 | # Days cut back on work due health | |
| 35 | B1PA5A | A1PA8A | BACA5A | RA1PA5A | Reason cut back on work (phys, me | Physical, mental or both | |
| 36 | B1PA5BA | A1PA8BA | BACA5BA | RA1PA5BA | Num days cut back due to phys hlth | # of Days physical | |
| 37 | B1PA5BB | A1PA8BB | BACA5BB | RA1PA5BB | Num days cut back due to ment hlth | # of Days mental | |
| 38 | B1PA5BC | A1PA8BC | BACA5BC | RA1PA5BC | Number days cut back due to ment | # of Days combination | |
| 39 | | A1PA9 | | | | Physical health at 16 | At M1 only |
| 40 | | A1PA10 | | | | Mental health at 16 | At M1 only |
| 41 | B1PA6A | | BACA6A | RA1PA6A | History of stroke | | Not at M1 |
| 42 | B1PA6B | | BACA6B | RA1PA6B | History of serious head injury | | Not at M1 |
| 43 | B1PA6C | | BACA6C | RA1PA6C | History of Parkinson disease | | Not at M1 |
| 44 | B1PA6D | | BACA6D | RA1PA6D | History of other neurological disorder | | Not at M1 |
| 45 | B1PA7 | A1PA11 | BACA7 | RA1PA7 | Heart trouble suspect/confirmed by | Heart problems ever | |
| 46 | B1PA7A | A1PA11A | BACA7A | RA1PA7A | Age doctor told you have heart prot | Age of heart problem | |
| 47 | B1PA7BA | A1PA11BA | BACA7BA | RA1PA7BA | Diagnosis - Heart attack | Heart attack | |
| 48 | B1PA7BB | A1PA11BB | BACA7BB | RA1PA7BB | Diagnosis - Angina | Angina | |
| 49 | B1PA7BC | A1PA11BC | BACA7BC | RA1PA7BC | Diagnosis - High blood pressure | High blood pressure | |
| 50 | B1PA7BD | A1PA11BD | BACA7BD | RA1PA7BD | Diagnosis - Valve disease | Valve disease/mitrovalve prolap | |
| 51 | B1PA7BE | A1PA11BE | BACA7BE | RA1PA7BE | Diagnosis - Hole in heart | Hole in heart/atrial septal dftc | |
| 52 | B1PA7BF | A1PA11BF | BACA7BF | RA1PA7BF | Diagnosis - Blocked artery | Blocked/closed artery/corony art | |
| 53 | B1PA7BG | A1PA11BG | BACA7BG | RA1PA7BG | Diagnosis - Irregular heartbeat | Irregular/fast heart beat/arrhyt | |
| 54 | B1PA7BH | A1PA11BH | BACA7BH | RA1PA7BH | Diagnosis - Hear murmur | Heart murmur | |
| 55 | B1PA7BI | A1PA11BI | BACA7BI | RA1PA7BI | Diagnosis - Heart failure | Heart failure/congestive heart | |
| 56 | B1PA7BJ | A1PA11BJ | BACA7BJ | RA1PA7BJ | Diagnosis - Other | Other heart trouble | |

V. Variable Formats

1. Utilize “Numeric” formats whenever possible; avoid using raw string variables that contain verbatim text. Open-ended responses, text, and verbatim data should be numerically coded where possible.
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 Barry Radler (bradler@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 Administrative 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 that are released publicly.**
5. Derived or constructed variables: if possible, any constructed or derived variables (e.g., scale 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 or DO NOT KNOW**
- 8 = REFUSED or MISSING**
- 9 = INAPP**

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

Ideally, all cells in MIDUS datasets should be populated with a value and empty cells should be avoided. The following values should be used to indicate different types of missing data or non-response to questions or data fields (a series of 9's can be appended to fit the maximum number of digits/integers for a particular variable):

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 directed to answer 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)

The Refresher and M3 introduced a new convention for coding SAQ non-responders in the Project 1 protocol. SAQ variables for those cases that did not return a completed SAQ will be coded "-1" with a value label "RESPONDENT DOES NOT HAVE SAQ DATA." Eventually all of the Project 1 SAQ variables will be coded and labeled in the same way.

OTHER NEGATIVE CODES (-2, -9, etc.)

It is increasingly clear that in many situations it is more efficient from the data manager's and the analyst's standpoint to use negative integers as non-response codes. MIDUS has begun adopting these in limited situations (such as for coding SAQ non-response). For the M1 Boston Longitudinal Cognitive data (M1P3), MIDUS began using "-2" to indicate that a case lacked longitudinal data and "-9" to indicate a case lacked baseline data. For the M3 Retention Early Warning data (M3P7), MIDUS used "-1" to indicate lack of SAQ data and "-2" to indicate lack of blood assay data.

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).