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

Conventions for Creating
MIDUS Refresher & MIDUS 3 Datasets

The MIDUS Administrative Core has developed conventions for creating MIDUS datasets for the MIDUS Refresher and MIDUS 3. 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, Milwaukee). 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 and M3. We have also included 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* format (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. Also end the file name with a datestamp, two examples of which follow:

Examples: Documentation/Instruments
      Refresher: **MR_P1_PHONE INSTRUMENT_5-8-12**
      MIDUS 3:   **M3_P1_DOCUMENTATION OF SCALES_20120508**

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

Examples: Datasets
      Refresher: **MR_P1_DATA_N2100_5-8-12**
      MIDUS 3:   **M3_P1_DATA_N5000_20120508**
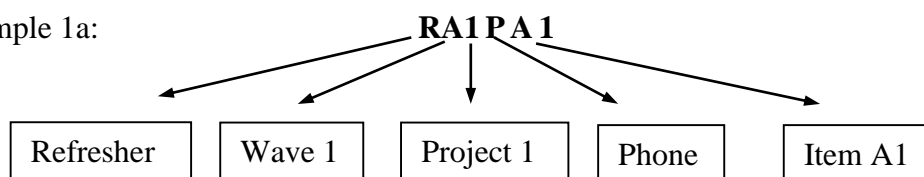
II. Variable Naming Conventions

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 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, 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. MIDUS 3 will follow the same conventions but will use 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.[1]
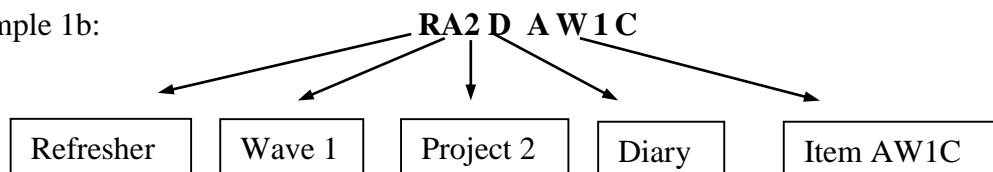
Examples: MIDUS Refresher

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.
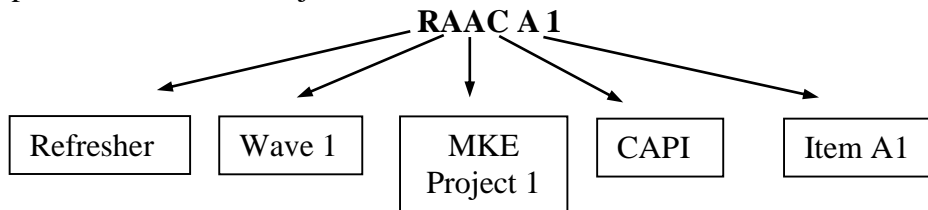
Example 1a: **RA1 P A 1**

| Refresher | Wave 1 | Project 1 | Phone | Item A1 |

Example 1b: **RA2 D A W 1 C**

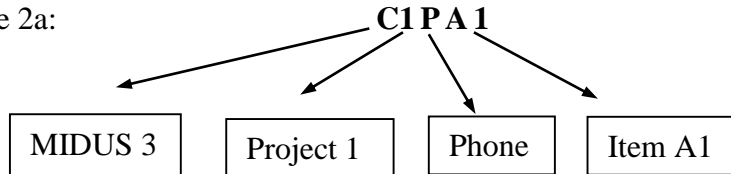| Refresher | Wave 1 | Project 2 | Diary | Item AW1C |

---

[1] 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.
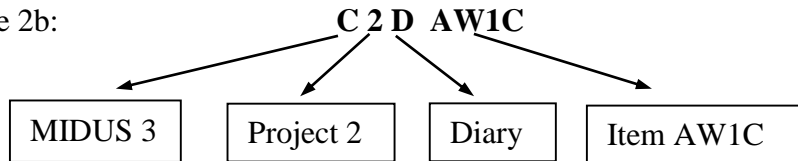
Example for Milwaukee Project 1:

RAAC A 1

```
Refresher    Wave 1    MKE        CAPI    Item A1
                       Project 1
```
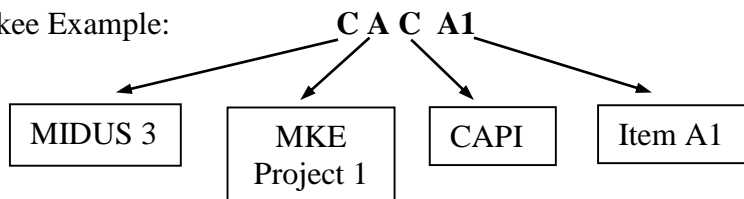
Examples: MIDUS 3

Example 2a:

C1 P A 1

```
MIDUS 3    Project 1    Phone    Item A1
```

Example 2b:

C 2 D AW1C

```
MIDUS 3    Project 2    Diary    Item AW1C
```

Milwaukee Example:

C A C A1

```
MIDUS 3    MKE        CAPI    Item A1
           Project 1
```

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 screen shot 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. Contact midus_help@aging.wisc.edu for more details if needed.

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)
> > The Refresher and M3 uses 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. Project 1 SAQ variables from M1 and M2 will eventually be coded in the same way.

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.