Guidelines for MIDJA Datasets

Naming, Coding and Formatting Conventions

We have developed guidelines for naming and coding conventions for the MIDJA data. Such conventions are necessary because we will need to merge records from multiple file types and sources across the survey and biomarker data collection. Without standardized conventions, this task will be unmanageable.

The attached pages provide specific guidelines for how variables will be named in MIDJA. We have also included coding conventions for "yes/no" variables as well as for "don't know" responses, "refused/missing" data, and "inapplicable" codes.

I.Variable Naming Conventions

A. Short Variable Names (SVNs): First 8 Characters (or less)

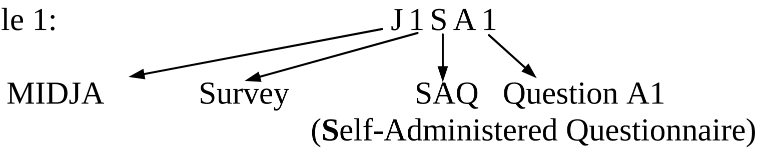
Notes:

- The first 8 characters (or less) must be unique for each variable.
- If you are doing analysis and/or programming code, using SVNs early in the process will prevent you from having to rename the variable names in your work.
- The first 3 characters of each variable name will identify the longitudinal wave in which the data were collected, the specific project, and the instrument used to collect the data. Characters 4 through 8 will identify the specific item or scale score variable that is represented by the variable name. Thus, "item identifiers" are limited to 5 characters.
- The letters/numbers in **bold** font, in the list below, represent the characters to be used in the variable name.

1st Character: **J** ("J" is being used to designate that the variable is part of the MIDJA data collection – all variables must start with this letter).
2nd Character: Identifies phase of data collection: i.e., 1= Survey 2=Biomarker) the second character of each phase).
3rd Character: Letter (Instrument Identifier: *type* of instrument, or, *name* of instrument; i.e., **S**urvey Questionnaire, **L**ab results, **Q**uestionnaire (biomarker) etc). Note that these Letters must be mutually exclusive ***within*** phases of data collection.

4th Character: Letter/Number (Item Identifier/Scale Name)
5th Character: Letter/Number (Item Identifier/Scale Name)
6th Character: Letter/Number (Item Identifier/Scale Name)
7th Character: Letter/Number (Item Identifier/Scale Name)
8th Character: Letter/Number (Item Identifier/Scale Name)

Example 1:

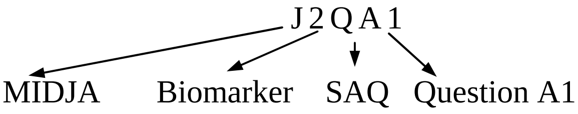


J 1 S A 1

MIDJA Survey SAQ Question A1

(**S**elf-Administered Questionnaire)

Example 1b:



J 2 Q A 1

MIDJA Biomarker SAQ Question A1

II. Variable Labeling Conventions

Variables labels will be no more than ***40 characters*** (considering the limitations across statistical applications) and we are using mixed case text for more sentence-like descriptions.

Example: **How is your physical health?**

III. Variable Formats

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

2. Variable formats should be precise - they should not exceed the maximum number of digits possible for a response. Thus, if a response will be 1-2 digits (i.e., something on a scale of 1-10) then the variable should be formatted as 2 digits.

3. Decimals: Limit to 2 places after the decimal point, unless important details of the data require more places be utilized.

4. Date/time formats: Dates or times recorded by *project staff* for administrative purposes (e.g. date & time blood samples were completed) can be submitted to the Core in the date and time formats specified in the established conventions for MIDJA (e.g., mm/dd/yyyy, or hh:mm:ss, etc.). Dates or times reported by *respondents*, especially in self-administered materials, cannot be submitted in date and time formats due to problems in applying missing value codes in SPSS. The individual components of a date or time must be recorded separately in their own variables. For example, dates will be broken into month, day and year variables, while times will be broken into hours, minutes and meridian (am/pm etc.)

IV. Value Labeling Conventions

A. Format

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

Example:
1 = YES
2 = NO
7 = DON'T KNOW

B. Coding Conventions for Yes/No Responses

YES = 1
NO = 2

C. Coding Conventions for “Non-response” (Don’t Know, Missing Data, and Inapplicable)

9’s will be used as place holders (i.e., if a variable is 3 digits, use 9 to fill in the first 2 places, then use 7, 8, and 9 for the final digit as displayed below).

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, 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.