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

Documentation of Coded Text Data in MIDUS 3 Project 7 (REW)

Background. This document describes how open-ended textual responses in the MIDUS 3 CATI and SAQ were transformed into categorical numeric codes. These codes are included in a stand-alone dataset (M3_P7_CodedTextData_N651_20210802.sav). This dataset contains only those cases (N=651) that contained text data in their responses.

Several items in the MIDUS 3 Project 7 survey ask participants to answer a question by providing nominal, categorical response options. For example, one item in the telephone interview asks participants about what types of cancer they had. Participants are provided with a variety of response options, such as "brain cancer," "lung cancer," "liver cancer," etc. With several exceptions, the instruments provide structured response options or metrics to capture answers to these types of questions. Building on the previous example, the phone instrument allows interviewers to record respondents" answers to the "type of cancer" question using 10 possible categories. These categories are labeled "breast cancer," "cervical cancer," "colon or rectal cancer," "lung cancer," "lymphoma or leukemia," "ovarian cancer," "prostate cancer," "skin cancer, melanoma," "uterine cancer," and "other___ (specify)."

The instrument's structured list of possible responses typically includes a response option of "other___(specify)." This response option allows the capture of responses not already provided by the instrument. During the telephone interview, when respondents provided a nominal response that did not belong to any of the provided categories, the interviewer would type the respondents" answer into the "other____ (specify)" field. For the SAQ, the respondent would write their response in the space provided under the "other___" field.

Note that some questions were true open-ended measures without a structured list of response options. Item AA1 in the phone interview, for example, asks respondents the specific event that made them aware that the recession had begun. Respondents were simply asked to volunteer their own responses to this question.

To make the M3 text data more usable, researchers at the University of Wisconsin's Institute on Aging categorized these text data. The remainder of this document explains the process by which text responses were coded.

The Process.

1. Preparing the Data for Coding. The verbatim text data from the M3 Project 7 phone interview and SAQ were delivered to the University of Wisconsin's Institute on Aging from the

University of Wisconsin Survey Center in Microsoft Excel files. These Excel files were cleaned to ensure that participants with "other specify" responses were represented in only one row of the spreadsheet, and that empty or duplicate rows were removed.

2. Coding the data. The categorization of text responses was accomplished using a text mining software program called IBM SPSS Text Analytics for Surveys. The extraction function of the SPSS-TA program served as a starting point for coding the responses. This extraction function is the default tool in the software program and identifies identical words, synonyms, and themes throughout the text data. Using this extraction tool, responses were coded into discrete categories by operating under the following guidelines:

(1) Each unit of information or theme from the text response field was coded once (i.e., was placed in one category only). However, in order to recognize that participants' text responses could include more than one unit of information, responses could be coded in more than one category for a single question. Multiple responses were recorded in multiple variables. For example, most participants responded to question AA1 with more than one theme such as "increase in gas prices and crashed stock market". This response contains two units of information: (1) increase in gas price, and (2) crashed stock market. This participant's responses were accordingly assigned two codes in two variables: one variable contained a code for their response of "increase in gas price" and another variable contained a code for their response of "crashed stock market."

(2) Categories were created such that their meanings would resemble the respondent's original answer as precisely as possible. For example, for the phone item F1 regarding participants' other Spanish origin, several respondents listed specific origin in Latin America, such as Honduran and Latino. Even though these responses refer to nationalities and ethnicities, separate categories for each of these origins were created (i.e., categories of "Honduran," "Latino").

(3) Categories of responses that did not exceed the 2% of the total valid responses were categorized as "Other."

At M2, content domain experts were consulted for help coding specific items. For example, the codes created types of cancer were reviewed by a biomedical researcher and a nurse practitioner. The codes created for items related to employment were reviewed by a sociologist with a background in labor force participation. The categories and codes derived from this process at M2 were used again for the M3 text data.

3. Packaging the newly coded information. Once all responses were coded in the SPSS-TA program, variables were exported to SPSS. These newly created "OS" variables were given variable names, variable labels, and value labels according to the following conventions:

(1) The variable names for the newly created "OS" variables parallel those of the original variables, but if there is more than one newly created "OS" variable for a single item, the newly created "OS" variables can be distinguished from each other by their having an "A," "B," "C," etc. at their end.

(2) All “OS” variables have variable labels that begin with “OS-”. Variables for open-ended questions have variable labels that start with “OE-”.

4. Coding for Phone Household Roster Non-normative Variables. In the M3 CAPI Household Roster section, there are questions that assess any non-normative conditions or developmental disabilities of the respondent’s child(ren).

When dealing with Household Roster variables, “entity numbers" must be taken into account. Each entity number represents an individual child of a respondent. For instance, if a respondent has two children, entity #1 and #2 refer to his/her two children, respectively. Entity numbers are represented in variable names and labels. For instance, if variable name ends with A1 (e.g. C7CCDTA1), “A” represents that this is the first response about non-normative conditions, and “1” represents that it is non-normative condition of respondent’s child #1. Similarly, if variable name ends with B3 (e.g. C7CCDTB3), “B” means that this is the second response regarding non-normative conditions, and “3” denotes that it is non-normative condition of respondent’s child #3. Variable labels also clarify the number of conditions for each child.

Many text responses to the non-normative and development disabilities questions were classified as “996: Some other condition (not non-normative)” because those responses were related to physical health problems or disease and were not strictly non-normative conditions, developmental disabilities, or mental health problems.

For additional information regarding M3 text data, please contact Barry Radler at bradler@wisc.edu.