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

Documentation of Coded Text Data in MIDUS 2 Project 1 (Survey)

Background. This document describes how open-ended textual responses in the MIDUS 2 CATI and SAQ were transformed into categorical numeric codes. These codes are included in a stand-alone dataset (M2_P1_CodedTextData_N4963_20201105.sav).

This document describes how open-ended text from questions in the MIDUS Project 1 interview and self-administered questionnaire (SAQ) are coded into categorical variables. Several items in the survey ask participants to answer a question by providing them with a list of response options. For example, the phone interview asks participants about what types of cancer (if any) they have had. Participants can choose from a list of 10 possible categories response options including "breast cancer," "cervical cancer," "colon or rectal cancer," "lung cancer," "lymphoma or leukemia," "ovarian cancer," "prostate cancer," "skin cancer, melanoma," "uterine cancer," but it also includes a response option of "other___(specify)." This response option allows the capturing of responses not already provided in the response option list. During the interview, the interviewer recorded these verbatim responses; in the SAQ, the respondent would write in their volunteered response in the space provided under the "other, please specify___" field. Note that some questions in both the interview and SAQ are simply "open-ended" (OE) questions and not closed-ended questions with an "other specify" (OS) response option.

To make the MIDUS text data more usable and to limit the release of potentially disclosing or identifying information that may be in the text data, the MIDUS Admin Core has developed a process to categorize and code such raw text data. The remainder of this document describes in detail the entire process of coding text survey data using SPSS Text Analysis for Surveys.

The Process.

1. **Preparing the Data for Coding.** The verbatim text data from the M2 Project 1 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, item F2 in the phone interview involves having the interviewer ask participants about their main ethnic background. One participant responded to this question by stating "Norwegian and Finnish," which the interviewer typed under the "other_____(specify)" field. This response contains two units of information: (1) that the participant listed Norway, and (2) that the participant listed Finland. This participant was therefore assigned two codes: one indicating their response of "Norwegian" and the other indicating their response of "Scandinavian." This participant was excluded from any additional categories (such as the category of "Western European").

(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 F2 regarding participants' main ethnic background, several respondents listed specific countries in Western Europe, such as Spain, Scotland, and Germany. Even though these responses refer to nationalities and not ethnicities, separate categories for each of these countries were created (i.e., categories of "Spain," "Scotland," and "Germany").

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

(4) After each participant's "other____ (specify)" responses were coded using these principles, the researchers created documents that summarized the common categories, giving each category a label and describing the types of responses included under each category. This information was then shared with other researchers with expertise in the area to which each item belonged. For example, the codes created for items related to 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.

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-”.

(3) With the exception of items C420 and C475 in the telephone interview, whose categories were assigned three-digit codes in the instrument’s structured list of responses, all “other____ (specify)” responses were assigned codes with the numbers “1110” or greater.

4. **The Product.** The result of this coding process is a standalone dataset including 142 variables and 4,963 cases: M2_P1_CodedTextData_N4963_20201105.sav. These data can be merged with any MIDUS Core sample dataset using “M2ID” as the keyed variable.