ICPSR | INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH

ICPSR 4652

Midlife in the United States (MIDUS 2), 2004-2006

Documentation of Coding Text Responses in MIDUS 2

Inter-university Consortium for
Political and Social Research
P.O. Box 1248
Ann Arbor, Michigan 48106
www.icpsr.umich.edu

Midlife in the United States (MIDUS 2), 2004-2006

Carol Ryff
University of Wisconsin-Madison

David M. Almeida
Pennsylvania State University

John S. Ayanian
Harvard University

Deborah S. Carr
University of Wisconsin-Madison

Paul D. Cleary
Harvard University

Christopher Coe
University of Wisconsin-Madison

Richard Davidson
University of Wisconsin-Madison

Robert F. Krueger
University of Minnesota

Marge E Lachman
Brandeis University

Nadine F. Marks
University of Wisconsin-Madison

Daniel K. Mroczek
Purdue University

Teresa Seeman
University of California-Los Angeles

Marsha Mailick Seltzer
University of Wisconsin-Madison

Burton H. Singer
Princeton University

Richard P. Sloan
Columbia University

Patricia A. Tun
Brandeis University

Maxine Weinstein
Georgetown University

David Williams
University of Michigan

Terms of Use

The terms of use for this study can be found at:
http://www.icpsr.umich.edu/cocoon/ICPSR/TERMS/4652.xml

Information about Copyrighted Content

Some instruments administered as part of this study may contain in whole or substantially in part contents from copyrighted instruments. Reproductions of the instruments are provided as documentation for the analysis of the data associated with this collection. Restrictions on "fair use" apply to all copyrighted content. More information about the reproduction of copyrighted works by educators and librarians is available from the United States Copyright Office.

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

Documentation of Coding Text Responses in MIDUS 2

Background. Several items in the MIDUS 2 Project 1 survey ask participants to answer a question by providing nominal, multi-categorical responses. Nominal, multi-categorical responses are those that have no implicit quantitative value and that have more than two possible substantively meaningful responses (i.e., their content involves more categories than "yes or no" or "female or male"). For example, one item in the telephone interview asks participants about what types of cancer they had. Participants could respond to this item with a variety of nominal, multi-categorical responses, such as "brain cancer," "lung cancer," "liver cancer," etc.

With several exceptions, the instruments contain structured responses to capture the possible nominal responses that participants might have provided in answering these types of questions. For example, building on the previous example, the phone instrument allows interviewers to key in respondents' answer to the "type of cancer" question under 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 to items with nominal, multi-categorical responses typically include a response option of "other___(specify)." This response option allows the capturing of responses that do not fit under the responses already provided by the instrument's list. In the telephone interview, when respondents provided a nominal response that seemed not to belong to any of the other categories, the interviewer would type the respondents' answer into the "other____ (specify)" field. In the self-administered questionnaire (SAQ), the respondent would write in their unique response in the space provided under the "other___" field on the instrument. Note that some questions were simply open-ended responses and not "other specify" categories. For example, item A23 in the SAQ asks respondents what diagnosis their doctor gave for their chronic pain. For this item, respondents were simply asked to write in the diagnosis.

To make the MIDUS 2 data more readily usable, researchers at the University of Wisconsin's Institute on Aging categorized such text data. The text below documents the process by which the task of coding text responses was conducted.

The Process.

1. Preparing the Data for Coding.

The verbatim text data from the MIDUS 2 Project 1 phone interview and SAQ were delivered to the University of Wisconsin's Institute on Aging from the Survey Center in Microsoft Access files. A Microsoft Excel file was created for each variable with textual data. These Excel files were cleaned before transporting them into the SPSS Text Analysis Program. Cleaning ensured that participants with "other specify" responses were represented in only one row of the spreadsheet. Because the Survey Center double data entered text from the SAQ (as a data quality check), it was necessary to eliminate many duplicate responses in these files.

In some cases, the text for a participant in one row of the spreadsheet did not match the text for that same participant in another row of the spreadsheet. Several rules were applied to guide decisions regarding which information to retain:

(1) If the rows contained responses that were identical in meaning, but one row had a response listed with misspellings and the other did not, the row with the more proper spelling was retained.

(2) If the rows contained responses that were identical in meaning, but one row included a response with more information than the other, the row with more information was retained.

(3) If the rows contained responses that were unique in meaning, but that did not contradict each other, information from the rows was merged together.

(4) If the responses listed in the rows contradicted with each other, a coin was flipped to decide which row of information to retain, and the coin flip was documented.

Once an Excel file was determined clean (i.e., when it was confirmed that the file listed each participant who indicated an "other specify" response in a single row), the file was then imported into the SPSS Text Analysis Program for coding.

2. Coding the data with SPSS Text Analysis.

The researchers allowed the program's extraction function to serve as a starting point for coding the responses. Using this extraction tool, in conjunction with the information from the instrument's list of structured responses, the researchers coded responses into discrete categories by operating under the following guiding principles.

(1) Each unit of information from the text response field was coded only once (i.e., was placed under one category only), but to recognize that participants' text responses could include more than one unit of information, respondents could be coded under more than one category for a single item. 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 "Scandinavia." This participant was excluded from any additional categories (such as the category of "Western European").

(2) An effort was made to create categories that would indicate those "other specify" responses that would fit into one of the existing categories on the instrument's structured list. For example, continuing with the example of item F2 regarding participants' main ethnic background, several respondents' information in the "other specify" field constitutes the text "American Indian" or "Native American." One of the structured responses under which interviewers could record participants' responses to this item is the category of "American Indian or Native American." To help analysts readily identify that these participants whose responses are listed under the "other specify" field could likely be merged with participants who are listed under the main "American Indian or Native American" response option, an "other specify" category label of "American Indian or Native American" was created.

(3) Categories were created such that their meanings would resemble the respondent's original answer as precisely as possible. For example, continuing with the example of phone item F2 regarding participants' main ethnic background, several respondents listed specific countries in Western Europe, such as Spain, Scotland, Germany, etc. Rather than creating one category labeled "Western Europe" to capture all these responses, separate categories for each of these countries were created (i.e., categories of "Spain," "Scotland," "Germany," etc.).

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 in Text Analysis program.**

Once all responses from the text fields were coded under substantively meaningful categories, new multi-categorical variables were created to capture this information. Some items with an "other specify" response field have only one newly created "OS" variable. This is the case when all respondents who used the "other specify" field only listed one unit of information in the "other specify" field. Other items with an "other specify" response field have several newly created "OS" variables. This is the case when any respondent who used the "other specify" field listed more than one unit of information in the "other specify" field.

These newly created "OS" variables were given short variable names, variable labels, values for each categorical response, and value labels by adopting several conventions:

(1) The short variable names for the newly created "OS" variables parallel those of the original variables with two exceptions:
    - All newly created "OS" variables have short variable names that end with the letter "O" (standing for "other specify").
    - 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 "O1," "O2," "O3," etc. at their end. Please note that although some OS responses are labeled as response #1, #2, #3, etc., *this order reflects the order by which respondents actually listed each unit of information*, nor their order of importance to the respondent.

(2) All newly created "OS" variables have variable labels that start with "OS-". Newly created variables capturing open-ended responses are given a variable label that starts 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 "110" or greater.

4. **Tips on working with items that have text responses.**

The newly created "OS" variables are intended to help analysts maximize the amount of substantively meaningful information collected from respondents. We encourage analysts to incorporate this information into their analyses when working with items that have multi-categorical responses. For the most part, analysts should be able to create variables that cleanly merge information from the original variables with that contained in the newly created "OS" variables. Here are two notes of caution to consider when doing so:

(1) When items allow respondents to list or check all that apply, many respondents have information under both the "other specify" field and a structured response. Take care in creating code that does not exclude one piece of information for the sake of another.

(2) When items ask respondents to list or check only one response, some respondents have information under both the "other specify" field and a structured response. To identify these cases, it is recommended to run a crosstabs analysis of respondents' answers to both the original item and all newly created "OS" variables. Then analysts can make an informed decision as to which piece(s) of information to retain for respondents with more than one response.

(3) Affected variables can be found in the dataset. Here is a list:

B1PA7BJA, B1PA28JA, B1PA31JA, B1PA32JA, B1PA33JA, B1PA34JA, B1PA35JA, B1PB3KA, B1PB5AO, B1P36AO, B1C420AO, B1C475AO, B1PD2O, B1PD4AO, B1PD8O, B1PE1GOS, B1PE1A7O, B1PF1BO, B1PF2AO, B1PF3O, B1PF7AO, B1PF8_AO, B1PF12AO, B1PC6AOS, B1PA22AO, B1PA23CJAO, B1SA14HAO, B1SA21FAO, B1SA23AO, B1SA43AO, B1SA50FAO, B1SA50GAO, B1SA51AO, B1SA55KAO, B1SA56SAO, B1SB8AAO, B1SB8BAO, B1SB8CAO, B1SB9AO, B1SB14DAO, B1SB17GAO, B1SC2AO, B1SN1AO, B1SP3JAO

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

Acknowledgements: UW graduate students Christina Chrouser Ahrens and Emily Greenfield contributed substantially to coding the M2 text responses and co-authored this document.