Considerations for Data Processing in Multi-mode Surveys

Erin Foster*, Lauren McNamara†, Kari Nysse-Carris‡

Tags: survey practice

Survey Practice

Vol. 3, Issue 6, 2010

Considerations for Data Processing in Multi-mode Surveys

The [Racial and Ethnic Approaches to Community Health across the U.S. \(REACH U.S.\) Risk Factor Survey](#) presented unique challenges to processing data collected from a multi-mode survey. REACH U.S., which is sponsored by the Centers for Disease Control and Prevention (CDC), monitors the progress and achievements of 28 local health interventions designed to eliminate health disparities among African Americans, Asians, Native Americans, and Hispanics. NORC at the University of Chicago conducts annual surveys in 28 communities using an innovative address-based sampling approach that combines multiple modes of data collection via telephone, mail, and in-person interviews.

In this paper, we discuss NORC's approach to processing multi-mode REACH U.S. data, paying particular attention to data processing procedures that occurred before, during, and after data collection. We also detail a metadata data processing framework that guided the creation of the final data file deliverables.

DATA PROCESSING CONSIDERATIONS PRIOR TO DATA COLLECTION

Data processing concerns, such as data capture and variable attributes, should always inform questionnaire design and programming, but these concerns are heightened when multiple modes of data collection will be employed for a single survey. Researchers should make every effort to ensure that variable construction is consistent across modes. Understandably, variable names displayed to the interviewer in computer-assisted telephone interview (CATI) and computer-assisted personal interview (CAPI) systems may be different than the variable name displayed to the respondent on a self-administered questionnaire (SAQ). However, the variable should be stored in the database under the same name and ask the same question across modes. For example, the question "Are you male or female?" should not be stored as SEX in one

---

* **Institution:** NORC at the University of Chicago

† **Institution:** NORC at the University of Chicago

‡ **Institution:** NORC at the University of Chicago

mode and GENDER in another; similarly, each variable should be asked with identical wording across modes.

Variable code frames also need to be consistent across modes. Response options in CATI and CAPI should match response options on an SAQ, and questions with identical response options across modes should be coded consistently across modes. Further, variables should be stored in the database with the same variable type (character or numeric) across modes. Merging data from multiple modes is simplest when combining variables of the same type.

Sometimes it is not possible to replicate variable structure – in question wording or how data are stored – across modes. For example, REACH U.S. asks respondents who have been diagnosed with diabetes two questions about how often the respondent and other health professionals check the respondent's feet for sores or irritations. It is likely that some diabetics have had their feet amputated and the questionnaire needs a response option to reflect that. In CATI and CAPI, interviewers are instructed to enter a special code (555) if the respondent mentions his or her feet have been amputated. After entering the 555 code at the first feet check question, CATI and CAPI are programmed to skip the second feet check question. On the SAQ, REACH U.S. included a space for the respondent to enter the number of times his or her feet were checked, and the option to check a box if feet were amputated. REACH U.S. then developed cleaning rules for what to do if a numeric response was given and the feet amputation box was checked, and what to do if the feet were marked as amputated at one of the two feet check questions but a numeric response was given at the other question. In the end, the SAQ's numeric answer and checkbox were reduced to one response that included a code to indicate the respondent's feet were amputated, and both questions about the feet check were consistent with each other.

Researchers should keep mode differences in mind throughout a multi-mode survey's design period, designing the SAQ to match the CATI and CAPI questionnaire as closely as possible and developing rules to structure SAQ data to match CATI and CAPI during data processing.

DATA PROCESSING CONSIDERATIONS DURING DATA COLLECTION

Researchers routinely incorporate data quality checks and edits into CATI and CAPI instruments, and it is rather straightforward to ensure that two computer-assisted questionnaires are identical with regard to programmed checks and edits. The structure of a hardcopy SAQ is innately different. CATI and CAPI questionnaires are programmed to follow skip patterns and trained interviewers assist respondents. These resources do not exist for SAQ, resulting in the need for very clear instructions and post data collection editing to resolve outliers, missing values, and inconsistencies. The REACH U.S. CATI and CAPI instruments are programmed so that only female respondents are asked about hysterectomies. While there is an instruction for only females to answer

that question on the SAQ, nothing prevents a male from answering “no.” As a result, post data capture edits – such as blanking out the response or coding “not in universe” – are required to correct erroneous responses to these questions by males.

DATA PROCESSING CONSIDERATIONS POST DATA COLLECTION

One of the greatest challenges to processing data from a multi-mode study is that duplicate interviews must be identified and reconciled. Duplicate interviews exist in a multi-mode study when a respondent inadvertently completes the interview in more than one mode. In REACH U.S., cases begin data collection in CATI or SAQ, depending if a telephone number is available for the sampled address. Cases that do not complete in CATI are then transferred to SAQ. Finally, a subsample of non-responding cases is randomly selected for in-person interviewing.

Due to the timing of interview completion and transfer to another mode, duplicate interviews can occur across modes. For example, a case may be transferred to CAPI while a returned SAQ travels through the mail. To reconcile these duplicate interviews, one option is to establish a priority order – for example, CATI first, then CAPI and finally SAQ. CATI and CAPI are prioritized due to the advantages of skip pattern programming and assisted interviewing.

A METADATA PROCESSING APPROACH FOR MULTI-MODE STUDIES

The multi-mode structure of REACH U.S. results in three input datasets for data processing. As can be seen in Chart 1, the input datasets are combined and output as a single, uniformly-formatted dataset. Processing programs reference a Microsoft Access® table – the “metadata dictionary” – to appropriately process the data.

Chart 1 Data Processing Flowchart.

Table 1 provides a simplified example of a metadata dictionary used to process multi-mode data. Each input variable is listed separately by mode because input details for the same variable sometimes differ across modes. For example, input variable length and type may differ. In the Table 1 example, the variables SEX and HIBPMEDCN are character variables of length 2 on CATI and CAPI, but are numeric variables of length 8 on SAQ.

Table 1 Example of the REACH U.S. Metadata Dictionary Design.

Id	Original name	Delivery name	Source	Original type	Original length	Master record	Delivery order	Deliver type	Delivery length	Delivery label	Format
1	CASEID	CASEID	CATI	CHAR	8	YES	1	NUM	8	Case identification number	
2	CASEID	CASEID	CAPI	CHAR	8						
3	CASEID	CASEID	SAQ	CHAR	8						
4	SEX	SEX	CATI	CHAR	2						
5	SEX	SEX	CAPI	CHAR	2	YES	2	NUM	3	Sex of respondent	SEXF
6	SEX	SEX	SAQ	NUM	8						
7	AGE	AGE	CATI	NUM	8	YES	3	NUM	3	Age of respondent	
8	AGE	AGE	CAPI	NUM	8						
9	AGE	AGE	SAQ	NUM	8						
10	HIBPMEDCN	BPMEDS	CATI	CHAR	2	YES	4	NUM	3	Medicated because of high blood pressure	YESNO
11	HIBPMEDCN	BPMEDS	CAPI	CHAR	2						
12	HIBPMEDCN	BPMEDS	SAQ	NUM	8						

Multiple records per variable are required for input processing, but cause complications for the output of a single variable. Creating a “master record” column to flag one row as the master output record solves this problem by defining variable order, final type, length and label once for each variable. While the output information is only listed on one row for each variable, its definitions are applied to all input variables so that the variable is output uniformly across modes. In addition, the master record flag provides the ability to filter the table to review the unique variables for output. Using filters on a single table rather than maintaining separate tables for input and output datasets simplifies version control.

The metadata dictionary contains additional columns to allow for renaming variables, keeping or dropping variables, setting variable value labels, and determining frequency display in code books. The delivery name column contains revised variable names for the final file as needed. For example, as can be seen in Table 1, a question regarding medication for high blood pressure, named HIBPMEDCN in production, was renamed to BPMEDS on the final data file using this column. All production variables are listed in the metadata dictionary, but many are not delivered, such as background variables to control skip patterns and text fills or variables used to derive a final variable for delivery. A variable is included on the final data file only if the variable is checked to be kept in the “keep column” of metadata. Variable value labels are assigned by a column that contains the format name. For example, SEX needs value labels of “male” and “female,” while BPMEDS needs value labels of “yes” and “no.” As can be seen in Table 1, formats of SEXF and YESNO are assigned. The format column is the key to link these variables with the Microsoft Access® table that contains the variable value labels. An example of the variable value format table is displayed in Table 2.

Table 2 Example of Variable Value Format Table.

Format	Code	Value label
SEXF	1	Male
SEXF	2	Female
SEXF	7	Don't know
SEXF	9	Refused
YESNO	1	Yes
YESNO	2	No
YESNO	7	Don't know
YESNO	9	Refused

Lastly, metadata contains a column to determine codebook display with assignments such as “suppress,” “frequency,” or “mean.” For CASEID, the variable is listed in the codebook, but frequencies are suppressed because they are not meaningful. Most variables have unweighted and weighted frequencies displayed. Some numeric variables may have mean and median values

displayed.

CONCLUSION

In summary, a multi-mode survey requires that data processing issues be considered throughout the life of a project. From questionnaire design to data collection to final data file creation, the use of multiple modes of data collection will require researchers to carefully consider ways in which to create a final file that accurately combines data from all modes.

We recommend that data processing requirements be carefully considered during the questionnaire design process. Designing consistent variables across modes in terms of name, question wording, code frame, and variable type and length will alleviate cleaning processes post data collection and increase the efficiency of the data delivery process. When question structure cannot be replicated across modes, each mode's variable should be designed so that they can be combined into a single variable during data processing. In addition, while SAQ does not have the advantage of CATI and CAPI programming and skip logic during data collection, the same skip logic should inform how the SAQ mode is cleaned to ensure consistency across mail, telephone and in-person modes.

Once data collection has concluded, the challenge is to combine multiple input data files into a single output data file. A metadata dictionary provides a detailed structure to a complex process. By including input variables from all modes with a "master record" flag to define the output variables, a single table can be used to define both input and output file information. In addition, the table can inform the structure of the final deliverable file in terms of the variables included, variable names, variable type and length, variable labels, code frame labels and codebook display.

Many of the above considerations are also important for surveys using only one mode of data collection, but they become more crucial when using multiple modes of data collection to ensure a complete and accurate final data file.