New Technologies in E-Census Data Collection Part 4: Developing Specifications for an Electronic Questionnaire

Select Topics in International Censuses[1]

Released November 2021

INTRODUCTION

Raw data contain errors that must be corrected before datasets are published or analyzed. In a census, this is done during data editing. Data editing is the process of detecting and correcting errors or logical inconsistencies. Data editing also helps to identify problems concerning missing or invalid responses. These problems are typically addressed using data imputation. Developing detailed computer assisted personal interviewing (CAPI) specifications is fundamental to preserving the original data and minimizing the amount of edited and imputed responses in a census dataset. This Select Topics in International Censuses (STIC) technical note provides guidance to National Statistical Offices (NSO) on how to develop specifications for an electronic questionnaire.

BACKGROUND

Census data editing has evolved over time. Initially, editing operations required large numbers of clerks to manually edit questionnaires. This task was labor intensive, difficult, and prone to error. The complex correlations between variables made it difficult for clerks to edit all the inconsistencies in data. Different clerks—and often the same clerk—would interpret the rules in different ways (United Nations, 2020).

The introduction of computers to data editing made this task less labor intensive, more accurate, and the application of editing rules more consistent. Computers also facilitated making alterations in datasets by automating changes to digitized records.

Yet another major advancement came with the introduction of electronic data collection. By using personal computers or tablets to enter respondents' answers, errors are minimized—in part—because CAPI allows for preprogramming the flow of questions based on previous responses, as well as validation checks of data entered in the field.

Paper census and survey questionnaires require that interviewers are trained to determine whether a question or section must be answered for a given person or household and compare responses with previous or subsequent responses to determine if the response is valid and logically consistent. This process is difficult, and errors are inevitable in a project as large as a census. For this reason, census data must be edited in the postenumeration stage of a census. Fortunately, postenumeration editing can be reduced in CAPI censuses by effectively using consistency checks—also known as data validation checks—which are dependent upon developing good edit specifications.

CAPI INSTRUMENT WORKFLOW IN AN E-CENSUS

Transitioning to CAPI from paper-based data collection impacts the timing of when the detailed edit specifications are finalized, as well as the composition of the experts required to develop them. Good specifications are required earlier in the CAPI timeline and the group working on them should include both subject matter specialists and computer programmers. Box 1 highlights these issues in more detail, drawing from a STIC that provides guidance to NSOs for planning an e-census (U.S. Census Bureau, 2016a).

[1] This technical note is part of a series on *Select Topics in International Censuses* (STIC), exploring matters of interest to the international statistical community. The U.S. Census Bureau helps countries improve their national statistical systems by engaging in capacity building to enhance statistical competencies in sustainable ways.

U.S. Department of Commerce
U.S. CENSUS BUREAU
census.gov

United States Agency for International Development
www.usaid.gov

United Nations Population Fund
www.unfpa.org

Box 1.

Planning for Mobile Data Capture

Census Timetable

In order to ensure the success of the census, the census timetable should be adjusted to fit the needs of implementing CAPI. Although the programming of the questionnaire cannot happen until the questionnaire specifications have been developed, it is a good idea to plan for the programming in the early stages of questionnaire development. With mobile data capture, previously separate processes may be integrated, or may need to be carried out earlier in the census life cycle. For example, data collection, capture, and editing can be done simultaneously in mobile data capture. However, generally, more time is needed to develop and test the application, set up the data transfer and processing systems, and procure, program, and test the mobile devices. If edit checks are to be added to the application, the edit programming must be completed before the enumeration begins, rather than programming those separately in the postenumeration stage. Further, more time should be allotted for training the census takers since the training must include the use of the mobile devices. Therefore, it is critical to determine all the steps needed to take place to set up a mobile data capture system and allow sufficient time prior to the enumeration in the census timetable.

Differences in Questionnaire Development Process

As with a paper questionnaire, developing an electronic questionnaire is an iterative process. It must be developed, tested, revised, and then tested again, repeating the cycle until the questionnaire works as intended. With an electronic questionnaire, the technical aspects of the electronic questionnaire application must be tested and revised, in addition to content. Further, after the subject matter specialists finalize the questionnaire content, specifications are needed to serve as a blueprint for programmers to design the application. Adequate documentation of the questionnaire instrument is also required. Not only does the lack of paper documentation make it difficult to discuss the questionnaire with various stakeholders, it also makes testing more time consuming and error prone since skip patterns are less obvious. In addition, an electronic questionnaire may contain additional features not included in a paper questionnaire, such as data validation and error messages. These features also must be specified so that programmers can design the application as intended. When designing an electronic questionnaire, the subject matter specialists may lose control over the wording, layout, and design of the instrument. Therefore, it is critical that the subject matter specialists work closely with the programmers to make sure that there is clear communication regarding the questionnaire content, layout and design, data validation, and other specifications. It is also important to have a good understanding about the timeline, ongoing changes to the questionnaire content, data security, and quality assurance.

Source: U.S. Census Bureau, 2016a.

Furthermore, since an electronic questionnaire allows for built-in data validation checks and does not have space limitations, questionnaire content and design issues are different from paper questionnaires. Box 2 highlights how data validation processes are built into the electronic questionnaire, drawing from a STIC on the developing of such questionnaires (U.S. Census Bureau, 2016b).

CONSISTENCY CHECKS

Errors in a census can result from either coverage or content. Coverage errors arise from omitting or duplicating people during enumeration, while content errors can result from: respondents giving faulty information, census takers asking a question incorrectly or misunderstanding the response, or census takers selecting the wrong response option.

Coverage errors are difficult to detect, though one approach is to list all the households and their members in an enumeration area prior to enumeration and then compare the information from the listing with the information collected during enumeration. Content errors are easier to detect. A well-developed CAPI application can greatly reduce content errors by applying consistency checks.

Consistency checks are tests programmed in a CAPI application to determine if data has any internal conflicts. Consistency checks help census takers identify errors while in the field, and allow them to fix problems immediately. For example, a consistency check should be programmed to prevent census takers from entering a higher number in the "Years of Residence" question than the number in response to the question on "Age."

In order to capture the complexity of data collection during enumeration, it is important that—in addition to data processors—subject matter specialists and field operation experts are involved in the development of consistency checks. A well-designed CAPI application should program enough checks to sufficiently prevent grossly inconsistent data while not prohibiting the census taker from entering data in a timely and unobtrusive manner. The presence of CAPI census consistency checks does not eliminate the need for post-capture edits, but well-designed checks will reduce the amount of post-capture editing necessary.

Types of Checks

There are two types of consistency checks, hard data validation checks and soft data validation checks. Hard checks prohibit the census taker from moving to the next question until a problem is resolved, while soft checks only warn the census taker that the data may not be correct. Hard checks must only be employed when there is certainty about whether the problem is an absolute error, for example, age at first marriage cannot be greater than age. The worst-case scenario with a hard check is that the check is programmed incorrectly and a census taker is not able to proceed with a correct response; the census taker may then have to input incorrect data in order to continue the interview.

When census takers encounter a soft check, they must review the response and determine whether to proceed with the interview or whether to fix the response. Soft checks are ideal for circumstances in which the check is generally valid but where there are occasional exceptions. For example, a soft check may be added to alert census takers that they are entering a response that is within the acceptable range, but uncommon, like when entering ages over 100 years. If in doubt, CAPI instrument developers should add a soft check that alerts the census taker, not a hard check.

Box 2.

Developing an Electronic Questionnaire

Data Validation

One advantage of using an electronic questionnaire is that it can validate the data as the enumerator enters the responses on the mobile device. To do this, data validation rules should be written by subject matter specialists in the questionnaire development stage so that they can be programmed into the application. Subject matter specialists are best suited for writing the validation rules because they have deep knowledge of the questions and possible responses.

Source: U.S. Census Bureau, 2016b.

Dependency of Checks

Generally, you want to check the validity of data as soon as possible so that the census taker and respondent can fix the error during the interview. Consistency checks can also be classified according to their timing in the questionnaire as unitary, in-occurrence,[2] and across-section.[3] Table 1 presents their dependency and timing.

[2] An occurrence refers to when the same data item appears multiple times in a questionnaire. For example, Relationship, Sex, and Age will be for every person in a household. In a household of five there will be five occurrences of each of these data items. In-occurrence means that we are to look within that specific occurrence. In this example, each person in the household would be an occurrence.

[3] A section is a group of related items, for example, education data, economic data, and fertility. A section can span all occurrences. For example, education data exist for all people over a specified age in a household; fertility data exist for all women (i.e., all occurrences of women) in a household. The CAPI application may be designed to require entering all the data for that section for all occurrences before moving to the next section.

Table 1.
Timing of Consistency Checks by Type of Check

Characteristic	Dependency	Timing	Example
Unitary	Depend only on the response to a given question. They are generally simple range checks.	Can occur as soon as the response is given.	Is the response in the valid list of codes?
In-occurrence	Depend on other responses given in a section for one occurrence.	Can occur once all dependent data has been entered, or after all data for the occurrence has been entered.	Is the person's age at first marriage less than or equal to the person's age?
Across-section	Depend on other responses given in a section across all occurrences.	Can occur once the dependent data for all occurrences of a section has been entered, or after all data for all occurrences has been entered.	Is one person specified as a head of household for the household?

Source: U.S. Census Bureau.

Unknown and Partially Known Responses

It is also important to determine whether a "missing" or "unknown" response should be allowed as a valid response in CAPI questionnaires. To enable some flexibility and to capture as much information as possible, some questions can allow "partially known" responses. For example, instead of an "Age unknown" category, it may be useful to have categories such as: "Age unknown but a child not yet in school," "Age unknown but a working-age adult," or "Age unknown but likely 50+." Partially known responses allow census takers to collect some data rather than leaving a response blank. Partially known responses can be assigned specific responses within the known range after data collection.

Scope of Checks

When developing consistency checks, it is important to notice that checks constrain the free input of data by census takers. In some instances, it is better to fix some problems postcapture rather than during data collection, as it is possible to add too many checks to a program. It is important to do a comprehensive analysis of all consistency checks before a census. The two main opportunities to do so are during a questionnaire pretest and during the pilot census.

The results of a questionnaire pretest and pilot census allow you to:

- Analyze what checks were triggered the most.
- Check if there are any glaringly inconsistent data that should be checked in future versions of the CAPI instrument.
- Ask the census takers if the checks inhibited a smooth data collection process.
- Check how much data passed through soft checks and determine whether the soft check should exist at all, if the soft check limits should be modified, or if the check should be changed to a hard check.

Developing Edit Specifications

The editing of specifications documents the checks to be conducted by the CAPI census applications. Specifications documents serve four main purposes: (1) subject matter experts can convey to programmers and other involved staff the necessary checks to be included, (2) programmers can use the document as they design the instrument, (3) subject matter specialists can review the checks for accuracy and completeness, and (4) data users (postcapture) can understand the parameters that went into data collection.

At a minimum, edit specifications should include the elements described in Table 2 (refer to appendix for full Electronic Questionnaire Specifications Guide). Table 3 represents a complete set of specifications for a hypothetical "relationship to head of household" question.

Table 2.
Electronic Questionnaire Specifications

Specification	Description
Number	Question number that matches the paper questionnaire.
Variable	Specify the name of the variable for the CAPI data file (e.g., P01_NAME_1, P01_NAME_2, ...).
Label	Specify the phrase that will appear on the top of the tablet screen.
Question text	The exact wording of the question to be read out by the census taker.
Interviewer instructions	Any special instructions for the interviewer.
Fills	Specify how an item that will be preloaded will be written, e.g., [NAME].
Universe	Specify for whom or what does this question apply (e.g., residents 3 years and older).
Responses	Specify all the possible responses depending on the type of question.
Routing	Specify what question should be asked next and for whom. Make sure you include specific routing for answers such as "Other: specify."
Checks	For each question, specify any consistency or data completeness check that is required.
Error messages	Specify what message should be issued for a given consistency or range check.
Programmer instructions	Describe any special instructions to the programmers.
Help menu	If including a help menu in the program, the menu content will also need to be specified.
Change log	Summarize changes made. Include date and name of programmer making the changes.

Source: U.S. Census Bureau.

Table 3.
Sample Specifications for Relationship to Head of Household Question

Specification	Sample
Number	A03
Variable Name	A03_RELATIONSHIP
Label	Relationship to head of household.
Question Text	What is [NAME]'s relationship to the head of the household?
Fills	A01_NAME
Data Type	Numeric
Capture Type	Radio Button.
Field Width	2
Interviewer Instructions	This field should not be blank.
Universe	All household members listed.
Responses	00- Head 01- Spouse/Partner 02- Son/Daughter 03- Son/Daughter-in-law 04-Step Child 05- Grandchild 06-Parent 07- Father/Mother-in-law 08 -Grand Parent 09- Brother/Sister 10- Nephew/Niece 11- Other relative 12- Not Related
Routing	If error message appears, (E1) please go back to A01_NAME and put the correct head of household. Otherwise go to A04_SEX.
Consistency Checks	1. Each household must have exactly 1 head of household. 2. Check if the first person listed is the head of household.
Error Message	E1: The first person listed MUST always be the head of household.
Programmer Notes	Soft check.
Help Menu	1. Head of household is a person who is at least 12 years old and is considered by other members to be head of household. In case head of household is less than 10 years old, should be allowed, in those rare cases. 2. It is possible to have one household head with more than one spouse.
Change Log	Misspelling on Question Text field corrected on 3/25/2018 by John Doe. Consistency check added on 5/28/2019 by Dr. Jane Doe.

Source: U.S. Census Bureau.

CONCLUSION

The process of collecting data is prone to content errors. CAPI instruments are superior to paper because they allow for building in preprogrammed edits that improve the quality of data as it is being collected. The development of good specifications for a CAPI instrument requires the cooperation of subject matter specialists, field representatives, and programmers. Developing a specifications document for a CAPI instrument is of critical importance as it can save time and resources and serve as valuable documentation of the decisions made. The goal is to improve data quality at the source, and to reduce the amount of data editing and imputation needed postcollection.

REFERENCES

United Nations Statistics Division, "Handbook on the Management of Population and Housing Censuses, Revision 2," New York, 2016.

United Nations Statistics Division, "Principles and Recommendations for Population and Housing Censuses, Revision 3," United Nations Publications, New York, 2017.

United Nations Statistics Division, "Handbook on Population and Housing Census Editing, Revision 2," New York, 2020.

U.S. Census Bureau, "New Technologies in E-Census Data Collection Part 1: Planning for Mobile Data Capture, 2016a," <www.census.gov/library/working-papers/2016/demo/mobile-data-1.html>, accessed July 6, 2021.

U.S. Census Bureau, "New Technologies in E-Census Data Collection Part 2: Developing an Electronic Questionnaire, 2016b," <www.census.gov/library/working-papers/2016/demo/mobile-data-2.html>, accessed July 6, 2021.

U.S. Census Bureau, "New Technologies in E-Census Data Collection Part 3: Timeline Impacts, 2019," <www.census.gov/library/working-papers/2019/demo/new-technologies-data-collection-part3.html>, accessed July 6, 2021.

Appendix

Electronic Questionnaire Specifications

Specification	Description
Number	Question number that matches the paper questionnaire.
Variable	Specify the name of the variable for the CAPI data file (e.g., P01_NAME_1, P01_NAME_2,...).
Label	Specify the phrase that will appear on the top of the tablet screen.
Question text	The exact wording of the question to be read out by the census taker. If the paper questionnaire used a grid format, then a verbatim question may need to be written. (e.g., Please list the names of all the people who slept at the household on census night, starting with the head of household). Specify prefilled questions and wording variation (e.g., Where was [NAME] born?) (See Fills below).
Interviewer instructions	Any instructions for the interviewer. Specify to indicate they are not read to the respondent. Specify to use different font or font color to differentiate them from the questions. This is also the place to specify the exact wording to be used to probe for responses.
Fills	Specify how an item that will be preloaded will be written, e.g., [NAME]. This is how the fills are identified. Specify the input source (such as variable names, outside data source).
Universe	Specify for whom or what does this question apply (e.g., residents aged 3 years and older).
Responses	Specify all the possible responses that can be generated from the question for categorical responses. Specify the allowable range for open-ended numeric fields (range checks). Indicate whether the range check should be soft or hard check. Specify input format for dates, telephone numbers, etc. Specify whether only one response can be selected, or multiple responses are allowed. Specify the graphical control element for the response (e.g., radio button, check box, etc.). Specify the character limit (i.e. field width) for open-ended character responses. Indicate whether “don’t know,” “refused,” or empty responses are allowed. If you do not include blank as an option, CSPro will not allow the response and will not allow the census taker to go to the next question.
Routing	Specify what question should be asked next and for whom. Make sure you include specific routing for answers such as “Other: specify.”
Checks	For each question, specify any consistency or data completeness check that is required. Indicate the specific variables or responses that are involved in the check. Indicate whether the check should be a soft or hard check.
Error messages	Specify what message should be issued for a given consistency or range check. Include routing if check is violated. Provide means to continue the interview when no resolution is possible.
Programmer instructions	Describe any special instructions to the programmers. Specify looping requirements. Indicate the number of times the program should loop through a series of questions (such as the number of jobs to collect in a job history).
Help menu	If including a help menu in the program, the menu content will also need to be specified.
Change log	Summarize changes made. Include date and name of programmer making the changes.

Source: U.S. Census Bureau.

The Select Topics in International Censuses (STIC) series is published by International Programs in the U.S. Census Bureau’s Population Division. The United States Agency for International Development sponsors production of the STIC series, as well as the bilateral support to statistical organizations that inform authors’ expertise. The United Nations Population Fund collaborates on content and dissemination, ensuring that the STIC series reaches a wider audience.