Coding Write-In Responses in a Census

Select Topics in International Censuses[1]

Released June 2020

INTRODUCTION

All census instruments include questions that require respondents to provide write-in answers. Write-in fields allow respondents to answer in free text format rather than with predetermined response categories. Write-in responses empower respondents by allowing them to answer questions freely, unconstrained by preset answer categories. However, the inclusion of write-in fields in a questionnaire poses particular challenges for National Statistical Offices (NSOs) during data collection and processing.

This Select Topics in International Censuses (STIC) technical note provides NSOs with information on internationally recognized standards on automated coding of write-in answers in a census.

QUESTION FORMATS

Closed-ended questions are the most common type of items in census questionnaires. These questions include a list of fixed-choice response alternatives and usually instruct respondents to select one or more of them. In contrast, open-ended questions have no preexisting response categories and allow respondents to answer questions freely, in their own words. In paper census questionnaires designed to be scanned, open answers are typically written in "write-in" fields, whose width determines the maximum number of characters allowed in the answer.

There are both advantages and disadvantages associated with closed- and open-ended questions. Typically, closed-ended questions with a fixed list of responses are easier to answer and analyze—thus lowering the burden for both respondents and NSOs during data collection, processing, and analysis. However, preestablished response categories often miss the wide array of answers that respondents would have produced had the question been asked in an open format. Open-ended or write-in responses allow respondents to express themselves in their own words based on their full knowledge and understanding. This comes at a price, though, since answers to open-ended questions are harder to process and analyze for NSOs. Open-ended questions may also represent a significant source of bias, because even a properly designed question may be understood in different ways by different respondents.

Typically, researchers at NSOs use an open-ended question when the constraints of the close-ended question outweigh the inconveniences of the open-ended question. In other instances, questions must be asked in an open format, as explained below.

[1] This technical note is part of a series on *Select Topics in International Censuses,* 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

Table 1.
Economic Activity, Occupation, and Industry: Botswana Population and Housing Census 2011

ALL PERSONS 12 YEARS AND OVER					
ECONOMIC ACTIVITY				OCCUPATION	INDUSTRY
What has ... been doing mainly since Independence Day 2010? **Seasonal work** 01 Paid 02 Unpaid **Non-seasonal work** 03 Paid 04 unpaid **Other** 05 Job Seeker 06 House work 07 Student 08 Retired 09 Sick Other (specify)	Did ... do any type of work for pay, profit or home use for at least 1 hour in the past 7 days? 1 Yes (GO TO A22) 2 No [If no, has ... worked at own lands/cattle?]	Since ... was not working, what did he/she do? 1 Actively seeking work 2 House work 3 Student 4 Retired 5 Sick Other (specify) **[If female GO TO A25** **If male GO TO next person]**	What was ... working as during the past 7 days? 1 Employee - paid cash 2 Employee - Paid in kind 3 Self-employed (no employees) 4 Self-employed (with employees) 5 Unpaid family helper 6 Working at own lands/cattlepost	What type of work did ... do in the past 7 days? To be precise, what were main tasks and duties? *(Probe as necessary, use two or more words to describe the occupation)*	What was the main product, service or activity of ... place of work? (*Probe as necessary, use two or more words to describe the Industry)* **GO TO A25 for Female, else** GO TO next person
A19(2)	A20	A21	A22	A23(3)	A24(4)

Source: Statistics Botswana, 2011.

Open-Ended and Mixed-Format Questions

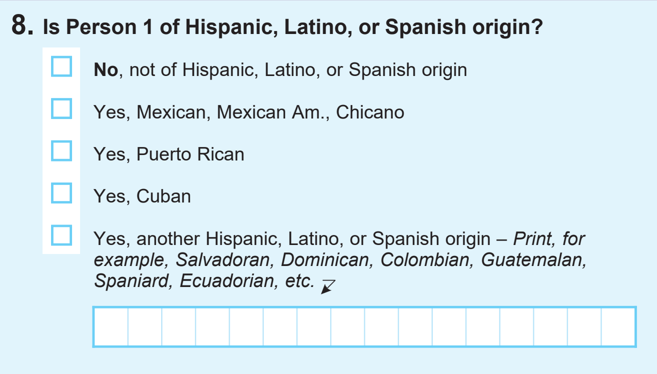
All census questionnaires have at least a one write-in question—name and surname of the residents of a household. Besides that, NSOs almost always allow write-in answers for some of their questions. For example, the responses to questions on industry and occupation are often open-ended. It is difficult to code or list the hundreds of occupations contained in an industry and occupation code list—especially in the case of paper questionnaires. Thus, NSOs often allow write-in responses when they want to capture more detail than what would be practical to display in a paper or digital questionnaire.

Table 1 provides an example of a census' Employment section (extracted from the 2011 Botswana Census). Notice that the Economic Activity questions are presented in a closed format, while the Occupation and Industry questions are asked in an open one.

Figure 1 presents an example of a question from the U.S. 2020 Census where four checkboxes and a dedicated

Figure 1.

Mixed-Format Question: U.S. 2020 Census Household Form



8. Is Person 1 of Hispanic, Latino, or Spanish origin?

□ **No**, not of Hispanic, Latino, or Spanish origin

□ Yes, Mexican, Mexican Am., Chicano

□ Yes, Puerto Rican

□ Yes, Cuban

□ Yes, another Hispanic, Latino, or Spanish origin – *Print, for example, Salvadoran, Dominican, Colombian, Guatemalan, Spaniard, Ecuadorian, etc.*

Source: U.S. Census Bureau, 2020 Census.

write-in field were combined to collect detailed information on Hispanic or Latino origin. The first checkbox is for respondents who are “Not of Hispanic, Latino, or Spanish origin.” The next three checkboxes are for the largest Hispanic origin groups—”Mexican,” “Puerto Rican,” and “Cuban.” The final checkbox of “Another Hispanic, Latino, or Spanish origin” is accompanied by a set of detailed examples and a dedicated write-in response area to collect data for all other groups in the Hispanic or Latino population. The combination of detailed checkboxes and a write-in area enables all respondents to report their Hispanic origin. This design represents a compromise between closed-ended and open-ended questions in that it is an open-ended question within a closed-ended format.

NSOs sometimes test different question formats until they find the one that best helps them collect the quality data they want, on time, and without straining their budgets. When implemented correctly, innovative question designs often pay off.

The “Name” Write-In Fields

The question on “name” is unique in that it can only be asked in an open-ended format, and is not intended as a means of retrieving relevant information for the census or survey. It is mainly used as an identifier. Names are often used to identify respondents within a household during data collection, and for matching purposes in Post-Enumeration Surveys (PES). Names are also used for indexing purposes in retrieval tools for census archives (see STIC on Census Data Archiving and Preservation).

Another particularity of “name” write-in responses in a census is that there is no need to code them. However, in order to properly use names as identifiers, spelling variations produced during data collection or capture should be addressed because it is often used for matching during the PES. For example, interviewers may miss the correct spelling of a name during personal or phone interviews. This is commonly found in names with alternative spellings or in phonetically similar names that do not share the same spelling. For example, Sean, Shaun, Shawn, and Shon are different spellings of the same name; and the last names Meier, Meyer, Maier, Mayer, Mair, and Mayr are all pronounced in a similar way.

Once census records have been made machine readable, names should be indexed to facilitate linking and retrieval of personal records. Box 1 provides an introduction to the Soundex system, one of the oldest and most popular phonetic indexing systems.

Box 1.

Phonetic Indexing and the Soundex System

The Soundex system developed by Russell and Odell in 1918 is often considered to be the first phonetical indexing system for surnames. By the 1930s, Soundex was already being used to index census data in the United States. The U.S. National Archives and Records Administration (NARA) still uses this system to locate records in historical census archives. Soundex is so flexible that its core principles can be adapted to different languages and for different uses. A Spanish version of Soundex is presented below.

The original Soundex code consists of one letter and three numbers. The letter is the first letter of the surname and the numbers are assigned according to the table below (if the last name is not long enough to produce three digits, trailing zeros are added). In this algorithm, vowels are ignored and consonants are grouped with similar consonants based on their phonetic place of articulation. Some additional rules apply for names with double letters, prefixes, letters side by side that have the same Soundex number, multi-word names, and others. Notice how last names Meier, Meyer, Maier, Mayer, Mair, and Mayr (mentioned above) all produce the same Soundex code, M600.

ORIGINAL (ENGLISH)

SOUNDEX CODING

Code	Letter
1	B, F, P, V
2	C, G, J, K, Q, S, X, Z
3	D, T
4	L
5	M, N
6	R
Discard	**A, E, H, I, O ,U, W, Y**

Source: Russell, 1918.

The table below corresponds to the “PhoneticSpanish” algorithm, a Soundex adaptation to the Spanish language. This table uses numbers 0–9 instead of 1–6; the letters “Y” and “H” are not discarded; and Spanish characters “LL,” “Ñ,” and “RR” are added to the algorithm. Special rules for some combination of characters apply. The “PhoneticSpanish” code consists of numbers only and it is not limited to four characters.

SPANISH PHONETIC CODING

Code	Letter
0	P
1	B, V
2	F, H
3	D, T
4	C, S, X, Z
5	L, LL, Y
6	M, N, Ñ
7	K, Q
8	G, J
9	R, RR
Discard	**A, E, I, O, U, W**

Source: Amón, Moreno & Echeverri, 2012.

CODING SYSTEMS

Once write-in responses have been digitized, they must be categorized into predetermined numbered classes, or coded. Write-in responses collected with digital means also have to be coded. There are three main ways to code: clerical coding, computer-assisted coding, and automated coding.

Clerical Coding

Clerical or manual coding is the least sophisticated coding system. Clerical coding is "coding by hand" and relies strictly on coders to complete the task aided by manuals and code lists only. It can be performed by interviewers in the field during data collection or by trained coders during data processing. In some instances, clerical coding is performed by respondents themselves when code lists are attached to the questionnaire.

Clerical coding of write-in responses in censuses is not a practical option due to the magnitude of responses. Clerical coding systems are often costly and time consuming, even in the smallest of countries.

Computer-Assisted Coding

In addition to clerical coding, write-in responses can be coded with a computer-assisted coding system, or CAC. In CAC, human coders work interactively with a computer while assigning codes. CAC is similar to clerical coding. However, in CAC, the coder has access to a number of computing resources like help screens, decision tables, auxiliary information, and others.

CAC systems are more advanced than clerical coding systems and often render savings in time and money for NSOs. CAC systems also play an important role in the evaluation and improvement of automated coding systems, as explained below.

Automated Coding

Automated coding systems are the most technologically advanced of these three coding systems. In automated systems, programmers run a program with detailed specifications that are used to classify (i.e., code) answers. Residual responses the system is unable to code are handled by human coders in CAC. Responses processed in CAC are then used iteratively to improve the system's ability to automatically classify further responses.

Automated coders reduce processing costs and ensure coding rules are applied consistently. However, even the most advanced automated systems require some degree of clerical coding. Clerical coding or CAC systems are necessary in the early stages of the development of automatic coders to ensure they work properly and to keep them up to date.

MODERN CODING SYSTEMS

Modern censuses require the use of a combination of automated and clerical coding to ensure that quality data are produced in a cost-effective way.

There are five basic parts or stages in an automated coding system (see Figure 2).

Stage 1: Dictionary Construction

Dictionaries for automatic coding systems are similar to code lists, except that dictionaries include as many variations of the same code as possible. A dictionary for a census automated coding system may contain hundreds of thousands of entries—depending on the size of the country. The dictionary should address common variations in spelling and language, and include variations caused by typos or scanning errors introduced by poor or blurry handwriting. For example, a dictionary for a hypothetical ancestry question would include entries such as: African, Afrcn, Africaine, Africam, Africano, Africsn, Afrika, i-african, Sfrican, etc. All of these entries represent typos, misspellings, and foreign words equivalent of the term "African" and should be coded in the same way. Dictionaries have to be built over time, as they will seldom capture any meaningful amount of potential responses initially.

Stage 2: Response Data Entry

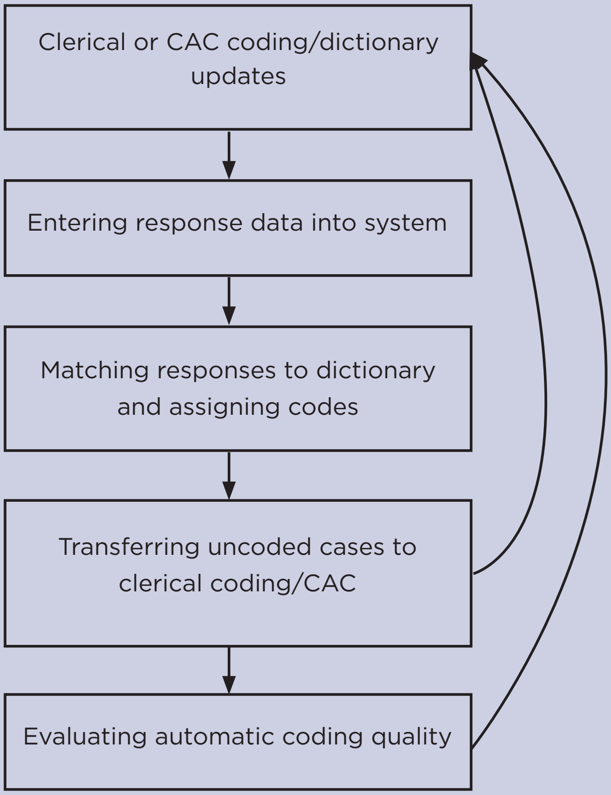
Digitized write-in data are entered into the system without corrections or edits. At this point, the only modifications to write-in responses allowed are the removal of invalid terms and characters.

Invalid terms include words or groups of words that do not constitute valid answers and do not change the meaning of the response when removed. For example, in a hypothetical ethnicity question, the answer "Half African and half European" should be coded as "African" and "European" separately. The word "half" and the conjunction "and" can be removed without altering the meaning of the intended response. However, in the answer "Not African, European," even though the term "not" is not a valid ethnicity answer, it cannot be removed without changing the meaning of the intended response. The term "not" should not be ignored but instead used to program an editing rule to ignore the term next to it, too. Thus, in this example only "European" should be coded as a valid response.

Invalid characters include commas, apostrophes, question or exclamation marks, mathematical symbols, and all other characters deemed invalid by the NSO. Removal of invalid characters only applies to responses retrieved from digitized paper questionnaires since only valid characters should be allowed in electronic collection modes like web-based or tablet applications. Often, removal of invalid characters is done during scanning or digitization.

Figure 2.

Automated Coding System


Clerical or CAC coding/dictionary updates
Entering response data into system
Matching responses to dictionary and assigning codes
Transferring uncoded cases to clerical coding/CAC
Evaluating automatic coding quality

Source: U.S. Census Bureau.

Stage 3: Response Matching

Once invalid terms and characters are removed from response entries, automatic coding systems try to match what remains of text strings to NSO-assigned codes using the dictionary. Ideally, all responses would have a match in the dictionary. However, in practice this is virtually impossible. In the response matching stage, responses with matching dictionary entries get a code assigned and those that the system is unable to match are flagged as residual responses.

Stage 4: Matching Residual Cases

Residual responses are transferred to CAC to be processed by trained coders who often specialize in one or more specific questions. Keep in mind that residual cases are expected to be more difficult to code and more uncommon than the average. For this reason, it is common for coders to have specific procedures to refer the most difficult cases to subject-matter experts for coding or adjudication. Once residual responses have been coded, the dictionary is updated with the new response-NSO code match. In this way, the system would be able to automatically match and code these responses in the future.

Stage 5: Quality Evaluation

The quality evaluation stage helps NSOs evaluate the quality of automated systems in order to improve them. To do this, a sample of successfully auto-coded responses are manually coded by the most senior coders. Then programmers can add new coding rules or make corrections to the automated system or amend the dictionary if necessary.

Box 2 summarizes the U.S. Census Bureau's 2019 Census Test coding operation for the Hispanic origin and race write-in answers.

Box 2.

The U.S. Census Bureau's Coding Operation and Referral Procedure

The Census Bureau has developed several automatic coding systems for different censuses, and surveys for different questions. However, the basic principles are the same. Coding procedures for the Hispanic origin and race questions in the 2019 Census Test are summarized below.

- In the 2019 Census Test, the coding operation consisted of two major phases: (1) automated coding and (2) residual coding.
- During the automated coding process, newly collected write-ins were coded by comparing them to records in the dictionary or "master file," which contains hundreds of thousands of previously coded write-in entries that have accumulated over several censuses and surveys and their assigned codes. A code was assigned when a match existed. Write-ins without a match required a trained coder to manually assign their proper codes.
- The residual coding process included three major activities:
    a. Production coding—Residual responses were manually coded.
    b. Verification coding—A sample of cases were coded a second time by a different coder.
    c. Adjudication—Discrepancies between the production coding and verification coding were resolved by a third, higher-rank coder or adjudicator.
- After residual responses were coded and passed the quality control measures, they were added to the master file to assist in the automated coding process by creating additional possible matches for future write-ins.

CONCLUSION

Developing efficient coding systems for write-in responses can be one of the most burdensome and challenging tasks in a census. Fortunately, with good planning and an efficient use of technology, the NSO's coding workload and budget can be reduced. A properly designed automatic coding system that incorporates some components of human coding would automate the process to the maximum extent possible without compromising quality.

REFERENCES

Amón, I., Moreno, F., and Echeverri, J., "Algoritmo fonético para detección de cadenas de texto duplicadas en el idioma español," *Revista Ingenierías, Universidad de Medellín*, 11(20), 127–138, 2012.

Bethlehem, J., "Applied Survey Methods: A Statistical Perspective," Wiley series in survey methodology, Hoboken, N.J., Wiley, 2009.

Campanelli, P., Thomson, K., Moon, N., and Staples, T., "The Quality of Occupational Coding in the United Kingdom," In Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N., and Trewin, D., eds., *Survey measurement and process quality,* pp. 437–453, New York, Wiley, 1997.

Groves, R., Fowler, F., Couper, M., Lepkowski, J., Singer, E., and Tourangeau, R., Survey methodology 2nd ed., Wiley series in survey methodology, Hoboken, N.J., Wiley, 2009.

Lyberg, L., and Kasprzyk, D., "Some aspects of post-survey processing," In Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N., and Trewin, D. eds., *Survey measurement and process quality,* pp. 353–370), New York, Wiley, 1997.

Rea, L., and Parker, R., "Designing and Conducting Survey Research: A Comprehensive Guide," fourth ed., San Francisco, CA, Jossey-Bass, a Wiley brand, 2014.

Russell, U.S. Patent No. 1261167, 1918.

Saris, W., and Gallhofer, I., "Design, Evaluation, and Analysis of Questionnaires for Survey Research," second ed., Wiley series in survey methodology, Hoboken, New Jersey, Wiley, 2014.

U.S. Census Bureau, American Community Survey, Design and Methodology, Washington, D.C., U.S. Department of Commerce, Economics and Statistics Administration, U.S. Census Bureau, 2009.

Wolf, C., Joye, D., Smith, T., and Fu, Y., eds., "The Sage Handbook of Survey Methodology," Los Angeles, Sage Publications, 2016.

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 informs authors' expertise. The United Nations Population Fund collaborates on content and dissemination, ensuring that the STIC series reaches a wider audience.