AN EVALUATION OF THE 1995 TEST CENSUS INTEGRATED COVERAGE MEASUREMENT (ICM) INTERVIEW: RESULTS FROM BEHAVIOR CODING

Nancy Bates, U.S. Bureau of the Census and Cynthia Good, Project Hope
Nancy Bates, U.S. Census Bureau, Room 3127-4, Washington, D.C. 20233-9150

KEY WORDS: CAPI, questionnaire design, reinterview

Introduction

The Integrated Coverage Measurement (ICM) process estimated the population of persons in the 1995 Test Census. The person interview component of the ICM was administered through a post-census CAPI interview. It was designed to gather an independent roster, conduct a thorough coverage interview, perform computer matching, resolve ICM/census roster discrepancies, and determine residency status. These competing goals, along with the complexities of a computer-assisted interview, presented numerous challenges for the questionnaire designers. The behavior coding explored in this paper provided valuable data from which to identify problems and begin the process of redesigning a simpler, more streamlined instrument.

This paper documents findings from a systematic review of a non-random sample of tape recorded ICM interviews conducted by a few interviewers at the Oakland and Louisiana test sites. These tape recordings served as the data for a questionnaire evaluation technique known as behavior coding, a questionnaire test methodology which systematically codes interviewer/respondent interactions. Despite some implementation problems, this method was useful in diagnosing problem question wordings, question order and the overall "flow" of the interview.

Results from the behavior coding revealed many problem areas with the 1995 ICM interview. Practically every question analyzed surpassed the "problem" cutoff level. These data indicate that interviewers frequently modified question wordings or failed to read questions. These findings were helpful in revising the ICM roster questions, re-ordering sections, revamping the approach for resolving ICM/Census roster discrepancies and rewriting questions to determine residency status.

Methodology

Field interviewers from the Oakland and Louisiana Census test sites tape recorded a sample of ICM interviews which served as the basis for the behavior coding. Our research plan specified the involvement of 10 interviewers from Oakland and 10 from Louisiana. We instructed ICM local supervisors to select these interviewers from the graduating class of the first ICM training module. We did not intend this procedure to follow a formally randomized selection process.

Once they began taping, interviewers were to keep tape recording until they completed 15 taped interviews. Given these guidelines, we expected to obtain approximately 150 tape recordings from each site, for a total of 300 recorded cases.

By the closeout of ICM interviewing, we had received 156 taped interviews from Louisiana but only 74 from Oakland. Eliminating unusable tapes, we were left with 122 Louisiana tapes and 64 Oakland tapes, far short of our requested 300 equally split between the two sites. The recordings from Louisiana were produced by 14 different interviewers, most of whom recorded between 10 and 15 interviews each. The Oakland tapes represented the work of only 8 different interviewers, 2 of whom produced over half of the tapes. The Oakland numbers obviously do not reflect the even case-load distribution that we had hoped for.

Since the assignment of interviewers and interviews was not random and the interviewer caseload was lopsided, the generalizability of the behavior coding results is compromised. This is particularly true in Oakland, where the majority of data come from just 2 interviewers.

Another limitation is the tape recording itself, which introduces unknown levels of bias into the research process. We suspect that interviewers may be more likely to follow the CAPI script, use flashcards, calendars, etc., when they know they are being tape recorded. Tape recording may also affect respondent behavior in unknown ways. Despite the limitations, we still felt that behavior coding was a worthwhile method for revising the ICM instrument. We believed that any problems uncovered during taped interviews would most likely be common to all interviewers.

Behavior Coding

Behavior coding is the systematic coding of the interactions between an interviewer and a respondent. Behavior coding is commonly used to assess whether interviewers have problems administering questions and whether respondents have difficulty comprehending questions, vocabulary, terms and concepts (Oksenberg, Cannell, and Kalton 1991; Morton-Williams and Sykes 1984; Marquis, Cannell and Robison 1971). This method is useful at indicating interviewer and/or respondent behaviors that may reflect problem questions, potential biases or inaccuracies in the data collection process. It's also a fairly inexpensive (although labor intensive) and relatively unobtrusive method compared to other pretest activities such as cognitive interviewing.

Two project staff members trained four experienced behavior coders to perform the coding. To behavior code the cases, coders listened to the tape recorded interview while simultaneously viewing the computerized trace file for the case. Trace files allowed the coders to "play back" the CAPI interview step-by-step, exactly as it occurred in the field.

This paper focuses on two components of the behavior coding scheme: question-asking codes and response codes. The first step, coding the initial question-asking behavior of the interviewer, is important because if many interviews show a deviation in wording on a particular question, it usually indicates that a question is poorly worded. Each question in the instrument was subjected to the behavior coding. Due to space limitations, this paper reports only on the results from the roster section of the interview.

The major categories for interviewer question asking behavior are as follows[1]:

Question Asking Codes:

Exact Wording or Slight Change - The interviewer asked the question exactly as written or with only slight modifications that did not change the meaning of the question.

Major Change in Question Wording - The interviewer administered the question with major changes to the scripted question wording that altered the intended meaning of the question (such as omitting key words, phrases, or dates or by paraphrasing).

Verification - The interviewer verified or repeated relevant information that the respondent had provided earlier, in place of asking a specific question.

Omission - The interviewer entirely omitted (answered without reading) an applicable question.

After coding the interviewer's presentation of a question, coders recorded the respondent's initial response to it. Coding respondent behavior is important for determining whether respondents are having difficulty understanding the meaning of questions and for identifying sensitive questions. Response codes are as follows:

Response Codes:

Adequate Answer - The respondent provided an adequate answer that met the objective of the question.

Inadequate Answer - The respondent provided an answer that did not meet the objective of the question and required additional probes to ascertain an adequate answer.

Break-in - The respondent interrupted with an answer before the interviewer finished reading the question.

Clarification - The respondent asked the interviewer to clarify the meaning of a particular question or concept, or asked for a repeat of the question.

Other Respondent Behavior - The respondent did something not covered by one of the other response codes (assumed non-verbal response, garbled recording, tape drop-out, etc.).

Whenever a major modification or inadequate answer occurred, coders recorded a brief note to indicate the specific modification or content of the inadequate answer.

Research indicates that behavior coding can be used to evaluate questions, but in order to do so, the coding must be reliable--that is, each coder must apply the same codes to the same behaviors. As an evaluation of the coders' grasp of the materials presented in training and to measure inter-coder reliability, we computed the reliability statistic kappa based on the same case coded individually by all four coders. We conducted inter-coder reliability tests at two different times, once immediately following training before full-scale coding began, and a second time, using a different case, about two-thirds of the way through production.

For each reliability test, we generated six kappa statistics in each category, one for each pair of coders[2].

For the first reliability test, the kappa for question asking codes ranged from 0.57 to 0.73, and for response codes ranged from 0.59 to 0.82. The overall percent agreement rate among the coders was 83 percent. The second reliability test yielded kappas for question asking codes ranging from 0.67 to 1.0 and response codes ranging from 0.38 to 0.89. The overall percent agreement rate among the coders was 87 percent. Since values of kappa above .75 are said to represent excellent agreement and values between .40 - .75 fair to good agreement, (Oksenberg, Cannell & Kalton, 1991), these results indicate that the reliability among our coders was within an acceptable range.

Following standard practice[3], we used 15% as a general guideline to indicate problem questions; that is, if 15% or more of the question readings had "problem" behaviors (e.g. a major change) then this indicated a significant level of the problem. We applied a more stringent cutoff in the case of question omissions (10% or higher was considered significant) because we felt this behavior was an obvious indicator of severe design problems (the question was perceived by interviewers as redundant, illogical, etc.).

Results

The first section of the ICM interview attempted to obtain the most accurate and thorough household listing possible. It began by asking for all persons living permanently or staying temporarily at the sample household on Census Day and was followed immediately by a battery of probes. Tables 1A and 1B illustrate selected questions from the roster section and summarize the interviewer and respondent behaviors.

Table 1A indicates that the roster section had both a high incidence of major wording modifications and relatively frequent question omissions. To begin with, the combination of the roster question (ROSTER) immediately followed by the order instruction (ORDER) was apparently not a very smooth start to the interview.

Interviewers made significant modifications to the roster question 20% of the time and over half the time for the order instruction. Common changes to the roster question included omission of the reference date and the phrase "staying here." A portion of this is probably due to the high respondent break-in rate (27% Table 1B). Interviewers frequently paraphrased the order instruction by shortening it or routinely omitting the last sentence. In 19% of the interviews, the order instruction was omitted altogether (Table 1A).

These interviewer behaviors may have contributed to some of the inadequate respondent answers at ORDER (see Table 1B) such as "it's just me and my little boy". Such answers do not meet the objectives of the question (a list of names). Other inadequate responses such as "I own my own home" resulted from the order instruction itself, which sidetracked respondents from their name-listing task.

In close to one-quarter of the cases, interviewers modified the introduction to the roster probes (INTRO) so that the meaning was changed (Table 1A). Most times it was paraphrased into something like "I have a few questions to make sure you didn't forget anybody," but some interviewers seemed to be warning the respondent of the probes to follow by making statements like "I know that this sounds redundant but they ask me to ask these questions ..." or "there's a few questions here, they may seem a bit redundant but they're designed for a reason." More unsettling is that in 11% of the interviews, the introduction was skipped altogether. We consider this a serious error since this statement serves as the only explanation of the critical roster review that follows.

Table 1A suggests that interviewers frequently modified the wordings of the first three roster probes (TEMPAWAY, ROOMMATE, CHILDREN). All three questions surpassed the 15% cutoff for major modifications. Some of these wording changes were due to reading only a partial list of the examples - e.g., "have I missed anyone temporarily away or on a business trip?" Other common errors were the omission of reference dates and the omission of clauses at the end of a probe - e.g., "in a general hospital", "live-in employee", or "child away at boarding school". Some interviewers tended to offer probes in a biased negative manner e.g., "no one off shore coming in for the weekend or anything like that?", "no roommates or foster children?" Perhaps the most common major modification was to collapse several probes into one by simply picking off one or two examples from each.

The frequency of omissions for the probes, ranged from 8% for the first, to double that amount for the last (16%). This may have been interviewers' response to respondent break-ins during earlier probes. It was apparent in some households (and in particular, one-person households) that the respondents perceived the probes as a redundant nuisance. This behavior helps illustrate a basic difficulty in the ICM interview. That

is, in order to uncover census omissions and determine erroneous inclusions, interviewers must apply intensive probing questions. These probes must be applied in each interview, but result in uncovering ICM/census discrepancies for only a small percentage. Consequently, interviewers are faced with a "needle-in-a-haystack" phenomena which requires patience and a good understanding of the survey's intent.

Roster Recommendations

Since the 1995 Test Census, the ICM CAPI instrument has undergone an extensive redesign program. Below are several changes to the roster section, many of which were the result of the behavior coding discussed here.

To improve the order instruction, we have moved the second part (ORDER) to become a separate question that follows the initial roster task in the revised instrument. It has also been shortened: "in whose name is this house/apt. owned or rented?" Interviewers then simply flag the line number of the appropriate person.

It is obvious from the interviewer behavior data that the extensive list of coverage probes was not administered as written. Sensing that the list is redundant, interviewers routinely shortened the questions, combined several questions into one, or simply omitted probes altogether. Interviewers also sometimes failed to provide an adequate context for the probes by omitting the introduction over 10 percent of the time. Since the quality of the independent roster is perhaps the most important component of the ICM interview, we recommended a completely new rostering technique for the next ICM instrument.

A new rostering alternative has been designed as part of the next ICM test cycle. The new approach guides respondents through the cognitive task of reconstructing their Census Day household roster using a process that differs from what they used in the census and also acknowledges the substantial time lag since Census Day. The approach first inquires about people who stayed at the sample unit on the night before the ICM interview. For each of these individuals, the interviewer determines whether the person also stayed in the unit on Census Day. Next, the respondent is asked a series of cues to aid recall of additional persons staying at the unit on Census Day who were not staying there the previous night (for example, persons who have moved away or persons away temporarily).

This approach (known as the retrospective approach) was designed to re-create Census Day rosters by first using information most accessible in memory (who was there last night) and then methodically working backwards (Biemer 1995; Sudman, Bradburn and Schwarz 1996). The resulting set of roster probes are shorter and more content-varied than the follow-up questions found in the '95 instrument. These efforts are meant to reduce the cycle of impatient interviewer interruptions and restructure the roster section by avoiding repetitive probes asked by rote.

The behavior coding presented in this paper served as a useful diagnostic toward the larger goal of building the next generation ICM instrument. During the next ICM test cycle, the revised instrument will undergo further rounds of coding plus other pretest methodologies such as interviewer debriefings, cognitive interviews, and usability tests. These activities should continue to improve the accuracy of data collected during the ICM interview, a critical component of the Census 2000.

Acknowledgements

The author thanks Kent Marquis of the Census Bureau for his helpful review of this paper.

References

Bates, N. and C. Good (1995). "The Integrated Coverage Measurement Evaluation of Tape Recorded Interviews with Behavior Coding: Project 2," DMD 1995 Census Test Results Memorandum Series, No. 18, December 15.

Bates, N. and M. Kindred-Town (1995). "The November Integrated Coverage Measurement (ICM) Test: Results from Behavior Coding of ICM Person Interviews," U.S. Bureau of the Census report, Center for Survey Methods Research, Statistical Research Division, April 10, 1995.

Biemer, P. (1995). "Draft Comments on ICM Evaluation Report Frames Related to Reconciliation Bias," Research Triangle Institute memorandum to H. Woltman, September 7, 1995.

Ellis, Y. (1994). "Categorical Data Analysis of Census Omissions," U.S. Bureau of the Census, DSSD 1990 REX Memorandum Series #PP-10, July 26, 1994.

Fowler, F. (1992). "How Unclear Terms Affect Survey Data," Public Opinion Quarterly, Vol. 56, 218-231.

Marquis, K., C. Cannell and S. Robison (1971). "Report on a Tape-Recording Analysis of Interviewer-Respondent Interaction in three Urban Areas. Chapters 10-12, in J.B. Lansing el al. (eds.) Working Papers on Survey Research in Poverty Areas, Ann Arbor: Institute for Social Research, 1971.

Morton-Williams, J. and W. Sykes (1984). "The Use of Interaction Coding and Follow-up Interviews to Investigate Comprehension of Survey Questions," Journal of Market Research Society, 26, 109-127.

Oksenberg, L., C. Cannel, and G. Kalton (1991). "New Strategies for Pretesting Survey Questions," Journal of Official Statistics, Vol. 7, No. 3. pp. 349-365.

Rolark, S. (1995). "Residence Rules for the 2000 Decennial Census," U.S. Bureau of the Census, memorandum for A.J. Norton from S. J. Rolark, October 20, 1995.

Sudman, S., N. Bradburn and N. Schwarz (1996). Thinking About Answers: The Application of Cognitive Processes to Survey Methodology. Jossey-Bass: San Francisco, 1996.

Sweet, E. (1994). "Roster Research Results from the Living Situation Survey," U.S. Bureau of the Census 1994 Annual Research Conference Proceedings, pp. 415-433, March, 1994.

Wellens, T. and E. Gerber (1996). "ICM Cognitive Evaluation," 1996 Census Test Memorandum Series Chapter IP-QD-1, June 28, 1996.

West, K. (1996). "ICM Evaluation Project 4: Response to Coverage Probes in the ICM Person Interview," DMD 1995 Census Test Results Memorandum No.21, January 29, 1996.

End Notes

1. The full list of respondent code categories also included multiple verifications and question collapsing; interviewer codes also included qualified answers and a "don't know" category. Since only the basic codes are presented here, row percentages in the tables may not always sum to 100.

2. Kappa statistics were calculated based on the full set of respondent and interviewer codes. Consequently, the estimated extent of agreement for the major categories of behavior is probably somewhat conservative.

3. The 15% cutoff is a standard index applied in other behavior coding research studies (see Oksenberg, Cannell, Kalton 1991; Fowler 1992).

This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.

Table 1A. Roster Questions - Interviewer Reading Behavior

(NOTE: Only the major codes are shown in the tables - row totals may not sum to 100%)

ROSTER SECTION	N	Exact/Slight Reading	Major Change to Q. Wording	Omitted Q. Entirely	Verified Answer
ROSTER	186	68%	20%	11%	0%
ORDER	182	29	52	19	0
INTRO	184	66	23	11	0
TEMPAWAY	185	58	32	8	2
ROOMMATE	185	67	22	9	2
CHILDREN	185	62	25	10	3
WORKWEEK	184	73	14	13	1
MOVED	184	73	10	15	1
NOPLACE	185	72	11	16	1

Table 1B*. Roster Questions - Respondent Behavior

(NOTE: Only the major codes are shown in the tables - row totals may not sum to 100%)

ROSTER SECTION	N	Adeq. Answer	Break-In	Clar-ification.	Inadequate Answer.	Other Behavior
ROSTER	15	47%	27%	13%	13%	0%
ORDER	134	75	1	8	15	2
TEMPAWAY	166	78	11	1	1	10
ROOMMATE	165	79	9	1	1	10
CHILDREN	162	78	10	1	1	9
WORKWEEK	161	83	4	2	1	10
MOVED	156	87	1	1	1	9
NOPLACE	156	87	3	1	1	9

* Responses at ROSTER reflect cases where ORDER was omitted and respondents answered at ROSTER. Because of interviewer omissions, the N's in table 1B are not the same as the N's in table 1A i.e., if a question reading was omitted, a respondent behavior was not coded.

Question Wordings for Roster Section:

ROSTER: What are the names of everyone who was living here permanently or staying here temporarily on March 4, 1995? **ORDER**: Please start with the name of the household member, or one of the household members, in whose name this house or apartment is rented, being bought or owned. If there is no such person, start with any adult household member. **INTRO**: We are trying to make sure that we count everyone in the census and count them at the right place. I am going to ask a few questions about people we sometimes miss. **TEMPAWAY**: Have I missed, anyone who usually lives here, but was temporarily away, spending the weekend with a parent, on a business trip, on vacation, or in a general hospital on March 4? **ROOMMATE**: Any housemate, roommate, foster child, roomer, boarder, or live-in employee? **CHILDREN**: Have I missed any young children, or babies born on or before March 4, or a child away at boarding school? **WORKWEEK**: Anyone staying here most of the week while working, even if that person has a residence somewhere else? **MOVED**: Have I missed anyone who lived here on March 4, but has since moved out? **NOPLACE**: Anyone who stayed here on March 4, who has no other place to stay?