Pascale, Joanne. 2016. "Behavior Coding Using Computer Assisted Audio Recording: Findings from a Pilot Test." *Survey Practice* 9 (2). https://doi.org/10.29115/SP-2016-0012.

ARTICLES

Behavior Coding Using Computer Assisted Audio Recording: Findings from a Pilot Test

Joanne Pascale[1]

[1] U.S. Census Bureau

Keywords: survey practice

https://doi.org/10.29115/SP-2016-0012

Survey Practice

Vol. 9, Issue 2, 2016

Behavior coding, a pretesting method that involves the systematic application of standardized codes to behaviors that interviewers and respondents display during the question/response process, has recently been enhanced by the use of Computer Audio Recorded Interviewing (CARI) system. Traditionally, interviewers used cassette tape recorders to capture a relatively small number of interviewers. CARI is built in to the CATI/CAPI instrument and allows for digital capture of all interviews (notwithstanding consent from the respondent). The Census Bureau piloted the use of CARI for evaluation of the 2010 American Community Survey (ACS) Content Test. This test was carried out to evaluate changes to existing questions that were designed to enhance data quality, and to evaluate alternative ways of asking about new topic areas. Interviews (or designated subsets of interviews) were recorded and coded in a total of 1,427 households. The coded interviews were conducted in both English and Spanish (1,092 and 335 cases, respectively) and they were recorded in CATI and CAPI modes (726 and 701 cases, respectively). This paper will provide highlights of the findings from behavior coding on a number of topic areas. It will also highlight the unique enhancements offered by CARI, such as: (1) dramatically increasing the volume and consequent range and diversity of recordings for more targeted analysis; (2) allowing researchers to listen to interviews during data collection in order to tailor behavior codes for the project; (3) allowing researchers to monitor the behavior coding operation in real time for quality assurance; (4) allowing the coder to see the actual screen display as the interviewer saw it when asking the questions; (5) allowing the coder to see the data keyed in to the instrument, enabling the identification and tabulation of keying errors; and (6) allowing behavior coding of both CATI and CAPI interviews, thus lessening the differences in effects of a centralized data collection facility versus a dispersed field staff not accustomed to regular monitoring and coaching.

Introduction

Behavior coding is a pretesting method where standardized codes are assigned to behaviors that interviewers and respondents display during the question/response process (Fowler and Cannell 1996). The method can be used to evaluate and improve questionnaires by, for example, helping identify survey questions that are problematic and identify aspects of interviewer training that could be strengthened. Until relatively recently, the utility of behavior coding was somewhat limited by sheer mechanics. Interviews were recorded on cassette tapes using a device attached to the telephone (for computer-assisted telephone interviewing [CATI]) or with an external recorder (for computer-assisted personal interviewing [CAPI]). Both methods proved onerous and limited the number of cases and the range and diversity of characteristics of cases that could be recorded. In 1999, RTI and the Census Bureau collaborated on a project to assess the feasibility of computer audio recorded interviewing (CARI), a system capable of capturing digital recordings directly on to CAPI

laptop computers (Biemer et al. 2000). In 2009, the system was adapted to include CATI recordings. The system also incorporated a software interface that allowed the coder to listen to the recording while viewing screen shots of the questions as displayed during the interview, and to enter behavior codes (including open-text notes).

This enhanced CARI system was first piloted in the 2010 American Community Survey (ACS) Content Test. The production ACS is first administered by mail; then a CATI interview is attempted for mail nonrespondents, followed by a CAPI interview with a portion of the CATI nonrespondents (Pascale et al. 2013). The 2010 content test mimicked this design, using experienced interviewers, and its purpose was to evaluate improvements to existing ACS questions and alternative versions of questions on new topic areas. The questions flagged for testing were inserted into the existing ACS questionnaire, and two versions of the instrument were developed – a test and a control. Behavior coding was one of several methods employed in the evaluation to aid the project sponsor in choosing between the test and control versions for production ACS. Nine topic areas and a total of 54 items were flagged for the behavior coding component of the evaluation. The main research question was whether the test or control version showed higher rates of standardized interviewer behavior and respondent behavior that was consistent with lower measurement error. (See Pascale et al. [2013], for a full report.)

The purpose of this paper is not to present the comprehensive set of results on the ACS Content Test per se, but to highlight some of the findings to demonstrate the advantages CARI brought to the behavior coding method. Many of these advantages were not necessarily unique to CARI, but the practical ease of recording (compared to traditional methods) dramatically increased the sheer volume of recordings that could be captured with minimal effort. This, in turn, greatly enriched the range of possibilities for analysis. Quotas of cases with specified characteristics – including relatively rare events such as receipt of public assistance – were identified a priori for recording. They were then coded in sufficient numbers to generate standard errors that could be used to evaluate test/control comparisons. Post-data collection, households and people with characteristics that emerged as being of interest for more in-depth analysis were identified for targeted behavior coding analysis. The topic areas of Food Stamps, public assistance and parental place of birth are highlighted below to illustrate these advantages.

There were also advantages unique to CARI. In terms of operations, research staff was able to listen to recordings while data collection was ongoing in order to develop tailored behavior codes before the survey was out of the field. Staff was also able to monitor coding operations in real time for quality assurance purposes and to conduct retraining as needed. And, because coders could see a

screen shot of the instrument and hear both interviewer and respondent audio at the same time, the coder could evaluate whether the interviewer correctly keyed in the respondent's answer.

Methods

The field period for the ACS Content Test was late August through mid-December 2010, with CATI cases followed by CAPI cases, and interviews in English and Spanish were conducted in each mode. To ensure an adequate number of recordings of each topic area without overtaxing the digital storage and transmission capacity, a quota of recordings was set for each topic area rather than recording the interviews in full. Interviews, or designated subsets of interviews, were recorded and coded for a total of 1,427 households. Of these, 77 percent (1,092) were conducted in English and the remainder in Spanish, and 51 percent were CATI and 49 percent CAPI. Eight bilingual telephone interviewers from the Tucson Telephone Center served as coders and were trained by staff from the Center for Survey Measurement (CSM). Training was held December 7–10, 2010, and coding operations were conducted from December 13, 2010, through March 6, 2011. Behavior coding data were then cleaned and processed by CSM staff.

For this study, the unit of analysis was a "turn" of speech for either interviewer or respondent. A turn begins when one person starts speaking and ends when the other person starts speaking. The starting point for development of the codes was a fairly standard set of behavior codes, which was adapted based on the analysis goals and by listening to recordings from the field. A measure of inter-rater reliability was calculated by assigning a subset of eight cases to all eight coders and then using the kappa statistic to measure the agreement across coders. According to Fleiss (1981), kappa scores can be categorized as follows: higher than 0.75 represent an excellent level of agreement, 0.40 to 0.75 represent a "good" to "fair" level of agreement, and scores below 0.40 indicate poor agreement. Overall, the kappa score for interviewer behavior codes was 0.502, and for respondent codes, the score was 0.463. One factor contributing to the relatively low reliability was that the recordings were sometimes out of sync with the item name and screen shot. To reduce file size and transmission time in the pilot, rather than make a continuous recording for a given case or topic area, recordings were made at the question-level. The recorder switched on when the interviewer entered an item screen and turned off when the interviewer moved off that screen. In many instances, this was problematic because interviewers moved on to the next screen before waiting for the answer, so the respondent's full response was cut off.

Results

Interviewer First-level behavior

Across all 54 items targeted for behavior coding, there were 20,352 administrations of questions, for an average of 377 administrations per question. See Table 1 for summary results of interviewer first-level behavior

codes for a subset of items. "Standardized" behavior indicates that the interviewer read the question as worded (or with a slight change) or correctly verified the question. Note that for some topic areas (e.g., computer devices), there is a test version of the question and a corresponding control version; the only difference was in placement in the questionnaire. For other topic areas (e.g., property income), there is not a one-to-one match of items, but a whole control series vs a whole test series. For the most part, for topic areas where a whole series of different questions was changed, the test version decomposed or clarified the original control question into simpler component parts.

Results show that overall, interviewers displayed standardized behavior 45 percent of the time, with no difference between test and control items as a whole. However, there was wide variation across topic areas and items. There were few test-control differences among items where the wording was the same across versions, as in the internet subscription type, computer device and parental place of birth items. For most of the topic areas where the test version was a decomposition of complex questions (property income, wages, veterans status/military service), rates of interviewer standardized behavior were higher in the test than the control version. For both Food Stamps and public assistance, there were large and significant differences – 73 vs 34 percent (control/test) and 44 vs 22 percent (control/test), respectively.

The test versions of both the Food Stamps and public assistance items were modified in an attempt to reduce underreporting. For Food Stamps, the program name had recently been changed to the Supplemental Nutrition Assistance Program (SNAP). The control version displayed this new program name in an optional interviewer instruction, while the test item embedded the new program name in the question itself (see Figure 1). The open-text notes on nonstandardized readings were categorized and quantified (see Table 2). Results show that in the control, 73 percent (row 1) read the question verbatim and that interviewers, as per instructions, never read the new SNAP program name from the optional text in the initial administration of the question. In the test version, only 34 percent read the question verbatim, but an additional 32 percent provided both the old and new program names (adding rows 2, 3, and 13), though they modified other parts of the question. Thus, in total, in 66 percent of the test cases, respondents were provided with the "Food Stamps" and "SNAP" stimuli, compared to no respondents receiving the SNAP stimuli in the control version. However, in the remainder of the test cases, both SNAP and "Supplemental Nutrition Assistance Program" were dropped – meaning that respondents did not receive any version of the new program name in these cases.

Unlike in Food Stamps, where the actual name of the program had changed, the test version of the public assistance item was modified to highlight certain aspects of the program that were suspected to be driving some of the underreporting: receipt on behalf of children and participation for as little as

Table 1 Interviewer first level frequency of standard behavior: control vs test.

Topic area	Item name	Control			Test			Diff % (T-C)	SE Diff	*p*-Value
		N	%	SE	n	%	SE			
Internet access	Overall	287	82%	0.023	443	75%	0.021	−6%	0.031	0.039
	AccessT				166	67%	0.037			
	InternetT				277	80%	0.024			
	SubscribeC	287	82%	0.023						
Internet	Overall	1,230	61%	0.014	1,045	63%	0.016	1%	0.021	0.486
subscription	Broad	159	62%	0.039	137	65%	0.041	3%	0.056	0.554
type	DSL	172	54%	0.039	149	58%	0.042	4%	0.057	0.497
	Dialup	164	68%	0.036	139	58%	0.041	−10%	0.054	0.074
	Fiberop	155	64%	0.039	137	66%	0.041	3%	0.056	0.649
	Modem	160	55%	0.040	135	60%	0.042	5%	0.058	0.388
	Othsvce	163	59%	0.039	133	60%	0.043	1%	0.058	0.827
	Satellite	162	68%	0.037	135	73%	0.038	5%	0.053	0.306
Computer devices	Overall	850	68%	0.016	826	68%	0.016	−1%	0.023	0.818
	Computer	279	72%	0.027	270	73%	0.027	1%	0.038	0.883
	Handheld	283	70%	0.027	273	68%	0.028	−1%	0.039	0.709
	Laptop	288	63%	0.029	283	62%	0.029	−1%	0.041	0.805
Food Stamps	FdStamps	288	73%	0.026	279	34%	0.028	−39%	0.039	0.000
Public assistance	PubAsst	913	44%	0.016	870	22%	0.014	−22%	0.022	0.000
Property income	Overall	1,037	27%	0.014	2,761	39%	0.009	12%	0.017	0.000
	IntrC	961	32%	0.054						
	IntrxC	76	27%	0.014						
	IntrT				911	25%	0.052			
	IntrxT				69	27%	0.015			
	RentT				887	12%	0.066			
	RentxT				25	50%	0.017			
	RoyaltyT				866	33%	0.333			
	RoyalxT				3	41%	0.017			
Wages	Overall	1,320	47%	0.014	2,093	51%	0.011	5%	0.018	0.010
	WagC	585	29%	0.019						
	WagxC	735	61%	0.018						
	EarnT				725	60%	0.018			
	EarntipsT				731	53%	0.019			
	TipstestT				40	15%	0.057			
	WagetestT				597	40%	0.020			
Parental	Ppobpa	924	14%	0.011	894	14%	0.012	0%	0.016	0.988
birth place	Ppobma	907	11%	0.010	877	10%	0.010	−1%	0.014	0.533
Veteran status/ military service	Overall	1,326	35%	0.013	1,310	55%	0.014	19%	0.019	0.000
	ActiveC	25	64%	0.098						

Topic area	Item name	Control			Test			Diff % (T-C)	SE Diff	*p*-Value
		N	%	SE	n	%	SE			
	MilC	585	59%	0.020						
	Vet1C	716	15%	0.013						
	ActiveT				22	82%	0.084			
	ReservesT				564	56%	0.021			
	TrainingT				23	70%	0.098			
	Vet1tT				701	52%	0.019			

Notes:

- If similar versions of the same question were asked in both test and control, a generic item name is shown and represents both test and control. Item names ending in "T" and C" were unique to test and control versions.
- "Standard" behaviors are exact reading/slight change and correctly verifying information provided earlier in the interview. "Non-standard" behaviors are major change, verifying in a non-neutral way, or skipping the question altogether. The numerators of the percent figures shown are standard behaviors only; the denominator is made up of all behaviors – standard, nonstandard and neutral/inaudible codes.

	CONTROL	TEST
Food Stamps	IN THE PAST 12 MONTHS, did anyone in this household receive Food Stamps or a Food Stamp benefit card? ♦ In some states the Food Stamps program may be known as the Supplemental Nutrition Assistance Program (SNAP) ♦ Do NOT include WIC or the National School Lunch Program	IN THE PAST 12 MONTHS, did you or any member of this household receive benefits from the Food Stamp Program or SNAP, the Supplemental Nutrition Assistance Program? Do NOT include WIC, the School Lunch Program, or assistance from food banks.
Public assistance	Did (<Name>/you) receive any public assistance or public welfare payments from the state or local welfare office DURING THE PAST 12 MONTHS?	Did [<Name>/you] receive any welfare payments or cash assistance from the state or local welfare office for [<Name>/yourself] or any children in this household DURING THE PAST 12 MONTHS, even if for only one month? Do NOT include benefits from food, energy, or rental assistance programs. ♦ See help screen for list of all State Welfare Programs.

Figure 1 Verbatim question wording for Food Stamps and public assistance (text shown in gray was optional, to be read at the interviewer's discretion).

one month (see Figure 1). Verbatim question-reading was 44 percent in the control and 22 percent in the test version. The most frequent type of change in the test version (24 percent of all administrations) was to stop reading after "...welfare office," meaning the question essentially reverted to the control version and included neither phrase intended to reduce underreporting. However, in 14 percent of test cases, interviewers made mention of the key changes (children and "at least one month") and another 11 percent mentioned children, even though they modified other parts of the question.

Table 2. Interviewer first level question reading, Food Stamps (control and test).

QUESTION AS READ (categories are mutually exclusive)		
Row	CONTROL TOTAL (n=288)	%
1	In the past 12 months, did anyone in this household receive Food Stamps or a Food Stamp benefit card? [verbatim]	73%
2	In the past 12 months, did anyone in this household receive Food Stamps?	6%
3	Receive Food Stamps?	4%
4	Interviewer read question as worded, but repeated part or all of the question after	4%
5	Did anyone in this household receive Food Stamps?	3%
6	Did anyone in this household receive Food Stamps, or a Food Stamp benefit card?	2%
7	In the past 12 months, did you receive Food Stamps?	1%
8	[Did you]* receive Food Stamps, or a Food Stamp benefit card?	1%
9	In the past 12 months, did you receive Food Stamps, or a Food Stamp benefit card?	1%
10	In the past 12 months, did anyone in this household receive Food Stamps, or a Food Stamp benefits?	1%
11	In the past 12 months, did you receive?	1%
12	In the past 12 months, did anyone in this household receive a Food Stamps card?	1%
13	Did anyone in this household receive Food Stamps, or Food Stamp benefits?	0%
14	In the past 12 months, did anyone in this household receive Food Stamp benefits?	0%
15	Skipped/Inaudible/Other	1%
Row	**TEST TOTAL (n=279)**	
1	In the past 12 months, did you or any member of this household receive benefits from the Food Stamp Program or SNAP, the Supplemental Nutrition Assistance Program? Do not include WIC, the School Lunch Program, or assistance from food banks. [verbatim]	34%
2	In the past 12 months, did you or any member of this household receive benefits from the Food Stamps Program or SNAP, the Supplemental Nutrition Assistance Program?	23%
3	In the past 12 months, did you or any member of this household receive benefits from the Food Stamps program or SNAP?	8%
4	In the past 12 months, did you or any member of this household receive benefits from the Food Stamps program?	8%
5	Did you receive Food Stamps?	4%
6	In the past 12 months, did you or any member of this household receive [some other wording]	2%
7	Did you receive [some other wording]	1%
8	In the past 12 months, did you receive [some other wording]	1%
9	In the past 12 months, did you or any member of this household receive benefits from Food Stamps?	1%
10	In the past 12 months, did you receive Food Stamps?	1%
11	In the past 12 months, did you or any member of this household receive benefits?	1%
12	In the past 12 months, did you receive benefits from the Food Stamps Program?	1%
13	In the past 12 months, did you receive benefits from the Food Stamps Program or SNAP?	1%
14	Did you or any member of this household receive Food Stamps?	0%
15	Skipped/Inaudible/Other	14%

The wording of the parental place of birth items was identical in test and control, as was the sequence (father then mother); the only difference was placement within the larger instrument. Both questions were asked at the

Parental place of birth	[Question wording was identical across Control and Test; the difference was in placement. Wording was also identical for paternal and maternal items (except the words "father" and "mother" as shown) and father's place of birth was asked before mother's place of birth]. In what country was [your/name's] [FATHER/MOTHER] born? Tell me the name of the country, or Puerto Rico, Guam, etc. ♦ Start typing the country or foreign place name and a look up coding box will appear. Select the appropriate country. ♦ If no country matches the respondent's answer, enter one of the following: ABROAD, AT SEA, or NOT LISTED.

Figure 2 Verbatim question Wording for parental place of birth (text shown in gray was optional, to be read at the interviewer's discretion).

Table 3 Parental place of birth interviewer first level frequencies by person number (control only).

Question/person	Standard		Nonstandard				Neutral
	Exact reading	Correct verify	Major change	Incorrect verify	Skip	Other	Inaudible
Paternal total (n=924)	10%	4%	68%	13%	3%	0%	2%
Person 1 (n=256)	18%	0%	73%	7%	1%	0%	0%
Person 2 (n=232)	7%	0%	77%	13%	1%	0%	0%
Person 3 (n=189)	7%	8%	61%	16%	6%	0%	1%
Person 4 (n=141)	6%	9%	62%	13%	4%	0%	5%
Person 5 (n=106)	7%	8%	55%	20%	5%	2%	4%
Maternal total (n=907)	5%	6%	62%	10%	14%	1%	4%
Person 1 (n=251)	8%	3%	76%	5%	6%	0%	3%
Person 2 (n=230)	3%	5%	65%	8%	15%	2%	3%
Person 3 (n=184)	5%	8%	48%	16%	18%	1%	3%
Person 4 (n=138)	4%	8%	49%	12%	20%	1%	8%
Person 5 (n=104)	4%	7%	52%	11%	17%	0%	10%

person-level about all household members (see Figure 2). The overall level of standardized behavior for these items was very low – 10–14 percent (see Table 1). In many households, the answer was the same for father and mother (that is, they were both born in the same country), and the answers were the same for all household members because they were all related. To investigate whether interviewers were reading the first administration of the question as worded and then abbreviating or skipping the question as they moved from one person to the next in the household, interviewer behavior by person number was examined. Table 3 shows a number of interesting results. For example, though the rate of exact readings did not drop as person number went up in a strictly linear pattern, there was a drop-off in exact readings after person 1. Furthermore, the overall rate of skips was 3 percent for the paternal version of the question but 14 percent for the maternal version.

Table 4 First level interviewer code frequencies by mode and language.

	Standard		Nonstandard				Neutral
	Exact reading	Correct verify	Major change	Incorrect verify	Skip	Other	Inaudible
TOTAL (n=20,426)	44%	1%	38%	4%	6%	0%	7%
By mode							
CAPI	40%	1%	39%	5%	9%	0%	7%
CATI	47%	1%	38%	3%	3%	0%	8%
By language							
English	53%	1%	34%	2%	3%	0%	6%
Spanish	36%	1%	42%	5%	7%	0%	8%

Comparisons by Mode and Language

Overall, the rate of standardized interviewer behavior was seven percentage points higher in CATI than in CAPI (see Table 4). The rate of major change was about the same across modes, but in CAPI, the rate of skips was 9 percent vs 3 percent in CATI, and the rate of incorrect verifications was also somewhat higher in CAPI than in CATI (five vs three percent). There were also differences by language; standardized interviewer behavior in English was 54 percent overall, vs 37 percent in Spanish.

Summary and Discussion

Table 1 suggests that decomposing complex questions into simpler parts results in interviewers reading the questions as worded more often than when several concepts are grouped together into one question. An examination of the nature of wording changes to both the Food Stamps and public assistance items indicates that interviewers did strive to deliver the key stimuli of the test question that was new/different, even if they did not read the question in its entirety. These kinds of results could be leveraged by linking to actual survey frequencies. For example, the Food Stamps items could be grouped into three categories: those where the question was read exactly as worded, those that were not read as worded but where the key stimuli (at least "Food Stamps" and "SNAP") were delivered, and those where neither key stimuli was delivered. Frequency of positive reports of Food Stamps would shed light on the implications of these changes.

Results from the parental place of birth questions suggest moving to a topic-based or household-based style of questioning for items where the answer is likely to be the same for all or most household members.

Results on mode are consistent with expectations. CATI interviewers are regularly monitored by their supervisors in a centralized facility. CAPI interviewers are not as accustomed to regular monitoring and feedback, which

could explain their higher rates of skipping and incorrectly verifying questions. The CARI system has the potential to "even the playing field" between CATI and CAPI and introduce CAPI interviewers to more frequent monitoring and more coaching from supervisory staff.

Many of the analyses conducted in this pilot are not unique to CARI; they are certainly possible using conventional behavior coding methods. However, CARI made it much more feasible to target specific topic areas and individual items to be recorded beforehand, and to flag specific characteristics of the interview after-the-fact for more in-depth analysis. Furthermore, this pilot only went as far as evaluating the interviewer-respondent interaction. The results could be leveraged for much more value by linking to the actual survey frequencies and conducting further analysis on the associations between characteristics of the interaction and the final survey estimates.

Due to unexpected technical difficulties, 20 percent of respondent first-level turns were coded as inaudible and most of these were driven by CAPI cases, which had an inaudible rate of 37 percent. This compromised the analysis of respondent behavior, as well as the data entry match. Overall, only one percent of cases were coded as a mismatch (that is, the answer keyed in by the interviewer did not match that provided by the respondent) and 75 percent were coded as a match. The remaining 24 percent of cases were "undetermined," and among these, the respondent's final answer was coded inaudible 83 percent of the time. Future use of CARI to evaluate CAPI interviews should include a pilot field test for audio quality of CAPI recordings. It is also recommended to record a continuous segment of an interview rather than at the question level.

REFERENCES

Biemer, P., D. Herget, J. Morton, and G. Willis. 2000. "The Feasibility of Monitoring Field Interview Performance Using Computer Audio Recorded Interviewing (CARI)." *Proceedings of the Joint Statistical Meetings of the American Statistical Association*. https://www.amstat.org/sections/srms/proceedings/papers/2000_183.pdf.

Fleiss, J. 1981. *Statistical Methods for Rates and Proportions*. New York: John Wiley & Sons.

Fowler, F., and C. Cannell. 1996. "Using Behavioral Coding to Identify Cognitive Problems with Survey Questions." In *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, edited by N. Schwarz and S. Sudman. San Francisco, CA: Jossey-Bass.