Using Computer Audio-Recorded Interviewing to Assess Interviewer Coding Error

Susan Mitchell, Matthew Strobl, Kristine Fahrney, Mai Nguyen, Barbara Bibb,
M. Rita Thissen, Wanda Stephenson
RTI International, 3040 Cornwallis Road, Research Triangle Park, North Carolina 27709-2194

Abstract

Computer-assisted audio recordings provide a new approach for detecting and correcting interviewer coding error. For questions with categorical and "other specify" responses, it is possible for the interviewer to misinterpret, abbreviate, or improperly key the respondent's answer. In this paper, we discuss the utility and effectiveness of computer audio-recorded interviewing (CARI) for detecting how accurately field interviewers capture responses to open-ended questions with precoded response lists that are not read to respondents. For this study, we evaluated coding accuracy by comparing the keyed data with the audio recording of what the respondent said. We examined four questions that differed in the length and complexity of the response lists, and the number of answers that were allowed (single- or multiple-response items). We present the type and rate of coding errors detected, and discuss implications for questionnaire design, interviewer training, and data quality.

Key Words: Computer audio-recorded interviewing, CARI, interviewer error, coding error, data capture error

1. Background

Interviewer-induced measurement error has long been cause for concern among survey researchers. A common approach to studying interviewer error is based on techniques for coding interviewer behavior. Behavior coding is used to yield counts of behaviors thought to produce errors, such as alteration in question wording, inappropriate or inadequate probing, and skipped questions (Groves, 1987). Although there is some evidence to the contrary (see review by Groves, 1989), researchers have found that open-ended questions are particularly sensitive to interviewer error (Fowler and Mangione, 1985; Mangione, Fowler, and Louis, 1989). This finding has been attributed both to failure to use probes appropriately and failure to accurately capture the respondent's answer.

Although most behavior coding studies have focused on interviewer mistakes during question administration, there has been little examination of interviewer error in recording responses. A few notable exceptions examine closed-ended questions. Rustemeyer (1977) compared interviewer entries with recorded verbal responses during mock interviews and found error rates ranging from 3.9% for experienced interviewers to 6.2% for newly trained interviewers. Kennedy, Lengacher, and Demerath (1990) used reports from monitors to detect an average keying error rate of .62% across four different telephone interviews. Dielman and Couper (1995) used tape-recorded personal interviews to detect an overall error rate of .095% for closed-ended questions. In a lone study examining interviewer effects in open-ended questions, Collins (1970) noted that interviewers' verbal idiosyncrasies (verbosity and word choice) can lead to measurement error when respondents' narrative responses are paraphrased and recorded.

The present study draws on the diverse features of this literature. We used recordings of four questionnaire items to examine interviewer error in capturing responses. The questions were asked in an open-ended manner; that is, interviewers did not read a response list to respondents, and respondents were free to answer in their own words. The interviewer's task was to interpret the response and find a suitable match on a precoded response list; if none of these applied, the interviewer entered the response into an "other specify" field.

Data for this study were captured using CARI, the latest in an evolution of technologies used to record interactions in survey research (Thissen et al., 2007). Recording was controlled by the laptops, and the sound recording was turned on and off without intervention by the interviewer. Although interviewers were aware that recordings would take place, they were not aware when CARI was turned on and which questions were recorded. CARI was used not only to study interviewer coding error but also to provide detailed and specific feedback to interviewers about their behavior, and to detect and deter interviewer falsification.

2. The Study of Community Family Life

This study is based on data from the Study of Community Family Life (SCFL). The SCFL is part of the Evaluation of the Community Healthy Marriage Initiatives, which is sponsored by the Administration for Children and Families and conducted by RTI International and the Urban Institute. The evaluation seeks to measure the community impacts of intervention programs designed to improve marital stability and satisfaction, and the well-being of families and children in low-income communities throughout the United States. The survey used computer-assisted personal interviewing to collect data from about 4,000 respondents in six sites: Dallas, TX; Milwaukee, WI; St. Louis, MO; Cleveland, OH; Fort Worth, TX; and Kansas City, MO. The survey instrument asked detailed questions about family composition, relationship status and quality, attitudes toward marriage and relationships, child well-being, and household and respondent characteristics. Data were collected from September 2007 through March 2008. Consent was obtained for CARI, and 92.6% of the respondents agreed to its use.

3. Assessing Interviewer Coding Error

As described earlier, four survey items were recorded for the purposes of this research. We selected the questions to represent a mix of types (single-response questions and multiple-response questions, with response lists of varying lengths) and complexities (eliciting a few words in response or many sentences). Each question was asked as an open-end, but instead of recording the response verbatim, interviewers coded the response using a list of precoded responses that were not read to respondents. Each item contained an "other specify" category for capturing responses that did not fit the precoded choices. The four items were asked at different points in the instrument to assess coding errors over the duration of the 45-minute interview.

From a list of completed interviews that was stratified by site and interviewer, we randomly selected 300 cases to review for each item, for a sample size of 1,200 cases. Prior to review, inaudible and empty files were removed from the sample, leaving a total of 1,083 audio recordings for analysis.

3.1 Hypotheses

The research team made four hypotheses regarding interviewer coding error prior to beginning the research. First, we hypothesized that coding errors would be impacted by the length of the response lists. A longer list with many precoded responses would be more difficult for interviewers to learn and search than a shorter list, increasing the potential for error. Second, we hypothesized that the number of responses allowed for each question would impact the error rate. The potential for making an error on a multiple-response item would be greater than a single-response item because the interviewer could fail to "record all mentions," an error type that does not apply to single-response questions. Third, we hypothesized that the complexity of the response list would impact the error rate. We defined "complexity" by the number of words each response label contained, whether the responses were narrow or broad (leading to errors of inclusion or exclusion), and whether respondents were likely to answer using the exact or similar words that appeared on the list. Finally, we hypothesized that questions created for the survey that had not been tested previously and used on other studies would be more prone to interviewer coding error. If we had no prior experience on which to base creation of a response list, and no knowledge of how respondents would answer the question, then we believed the response list could be insufficient or ill defined, leading to coding error.

3.2 Taxonomy of Interviewer Coding Errors

To perform the analysis, we developed a taxonomy of coding errors and used it to assign an error code to each audio file we reviewed. The error codes are shown in Table 1 below.

Table 1
Interviewer Coding Error Taxonomy

Error	Description
No error	The interviewer captured the respondent's answer correctly.
Selected wrong code	The interviewer selected the wrong code for the respondent's answer. This error type includes data entry errors and also errors of interpretation.
Recorded a listed response as an "other" response	The interviewer coded a response as "other" when it should have been coded as a listed response.
Recorded an "other" response as a listed response	The interviewer recorded a response using a listed response when it should have been coded as an "other" response.
Selected insufficient codes (applies only to multiple response items)	The interviewer did not code enough responses to capture the entirety of the respondent's answer.
Entered "other specify" text that does not match the response	The interviewer recorded text in the "other specify" field that does not match what the respondent said.
Did not capture the entirety of the "other specify" response	The interviewer recorded text in the "other specify" field that does not capture the response in its entirety.
Entered "other specify" text that was unintelligible/not codeable	The interviewer entered text into the "other specify" field that was unintelligible and not codeable.

3.3 Method of Assessment

To assess the error rates for each question, four reviewers who were knowledgeable about the survey listened to the CARI files, compared what they heard with what the interviewer had recorded, and assigned a code to indicate the presence and type of coding error. The reviewers worked in groups of two, with each person assigned two questions. To minimize burden on any one individual and to guard against coding bias, the two reviewers each listened to about half of a question's files. To ensure consistency among reviewers, about 10% of each reviewer's work was selected for double-coding by a second reviewer. Inconsistencies were noted and discussed, and corrections were made to the error coding as necessary. We calculated a final inter-coder reliability rate of 84.2%, indicating reasonable agreement among the reviewers. The discrepancies that remained were often due to audibility problems in the files, which left some responses open to interpretation.

4. FINDINGS

4.1 Overall Results

Across the four questions we reviewed, 85.5% were coded correctly, although this rate varied considerably by question. As shown in Table 2, the most frequently observed error, found in 4.3% of the files, was *selection of the wrong code*. This error occurred in three ways: when interviewers (1) miskeyed the number of the response, (2) misinterpreted the meaning of the code, and (3) misclassified a response based on insufficient information. The three error types are grouped together because we could not always distinguish among them with certainty. Our qualitative assessment, however, is that errors of interpretation and classification were more common than data entry errors. Errors of interpretation occurred when the response labels were not sufficiently descriptive for interviewers to code answers in the way questionnaire designers had intended. In addition, interviewers did not always share a common understanding of the meaning of a listed response, which led (albeit infrequently) to inconsistent coding across interviewers.

The second most frequent error, found in 3.2% of files, was *coded a listed response as "other."* This error occurred when interviewers had insufficient knowledge of the listed codes or did not take the time to sort through the codes to find a match. The third most frequent error, *selection of insufficient codes* occurred in roughly 2.7% of the files, although it was applicable to only two multiple-response items. Finally, errors related to recording answers in "other specify" text fields (*entered text that does not match the response, did not capture the entirety of the response,* and *entered text that was unintelligible*) occurred in 2.3% of the files.

The overall error rate, 14.5%, was somewhat higher than we expected based on prior studies of interviewer coding error. In whole or in part, however, it may be explained by the open-ended format of the questions in this study. This format requires respondents to formulate a response, interviewers to interpret that response, and then to sort through a list to find the correct match—all the while deciding if the response is better suited as an "other specify." These pressures are compounded by the need to maintain the pace of the interview and the respondent's interest while coding. In contrast, closed-ended questions require interviewers to simply record the number of the response from a list that is read to respondents. There is little or no opportunity to make many of the same errors associated with open-ended questions. Therefore, it is reasonable that the coding error rate for open-ended questions would be higher than the rate for closed-ended questions.

Table 2
Error Rates and Types across All Questions

Description of Error	Percentage (of 1,083 total files)*
No error	85.5%
Error rate	14.5%
Selected wrong code	4.3%
Recorded a listed response as an "other" response	3.2%
Selected insufficient codes (for multiple response items)	2.7%
Recorded an "other" response as a listed response	2.5%
Entered "other specify" text that does not match the response	1.5%
Did not capture entirety of the "other specify" response	0.7%
Entered "other specify" text that is unintelligible/not codeable	0.1%

*Percentages may total more than 100 because multiple codes were allowed.

4.2 Results by Question

This section discusses the results for each of the four recorded questions individually. The tables below show the complete text of each question and response list as they appeared in the survey instrument. Each question is labeled to show the number of listed responses, whether the item was single- or multiple-response, and the complexity of the response list (low, medium, or high). As mentioned, complexity was assigned based on a qualitative assessment of the list that considered (1) the number of words in the response labels, (2) whether the responses were broadly or narrowly defined, and (3) whether respondents were likely to respond with the same or similar words that appeared on the list (minimizing the need for interviewer interpretation or probing). Each table also shows the overall error rate for the question and a breakdown by error type.

4.2.1 How Is the Person (You Talk to Most about Marriage and Relationships) Related to You?

The first question, a single-response item, asked, "How is the person you talk to most about marriage or relationships related to you?" (see Figure 1). The response list contained 16 precoded responses of low complexity (e.g., mother, father, aunt). Despite having the longest response list of all the questions (16 responses), this item had the lowest error rate, 6.3%. However, compared with the other questions, the response list was easy to understand, the categories were well differentiated, and respondents were likely to answer in the same words. The most frequent error was *selection of the wrong code*, which occurred in 3.9% of the files. These errors represented both data entry errors and errors of interpretation or classification; for example, the following:

- Respondent said “a friend,” but the interviewer entered code 2 instead of code 12, most likely dropping the “1” unintentionally.
- Respondent said “coworker,” but the interviewer said back “friend?” and without disagreement from the respondent, coded 12 instead of 14.

Figure 1
How Is the Person Related to You?

Single-Response Item

16 Listed Responses

Complexity Rating: Low

How is (the person you talked with **most** about marriage and relationships in the past 6 months) related to you? For example, mother, father, friend, neighbor, someone you work with, a clergy person, or someone else?

MOTHER	1	OTHER RELATIVE	11
FATHER	2	FRIEND	12
DAUGHTER	3	NEIGHBOR	13
SON	4	COWORKER	14
SISTER	5	CHURCH/CLERGY	15
BROTHER	6	PSYCHOLOGIST/ PSYCHIATRIST/ OTHER COUNSELOR	16
MOTHER-IN-LAW	7	OTHER (SPECIFY)	17
FATHER-IN-LAW	8	DON’T KNOW	d
GRANDMOTHER	9	REFUSED	r
GRANDFATHER	10		

Error Types and Rates	Percentage (of 254 total files)*
No error	93.7%
Error rate	6.3%
Selected wrong code	3.9%
Recorded a listed response as an "other" response	1.2%
Recorded an "other" response as a listed response	0.8%
Did not capture the entirety of the “other specify” response	0.4%

*Percentages may total more than 100 because multiple codes were allowed.

4.2.2 What Is the Main Reason You Are Not Working Now?

The second question, “What is the main reason you are not working now?” is common in surveys measuring labor force participation. This single-response question contained 11 listed responses of medium complexity (see Figure 2). It had the second-lowest error rate, 11.5%.

The most common error for this item, found in 5.0% of the files, was *coded a listed response as an “other” response.* Much of the time, this error appeared to be caused by the interviewer not being fully familiar with the response list; for example, the following:

- The respondent said, “I’m taking care of my mother,” and the interviewer coded it as “other” when the response should have been coded as “3” (“taking care of home/family/children”).
- The respondent said, “I’m under a doctor’s care,” and the interviewer coded it as “other” when it should have been coded as a “1” (“ill or disabled and unable to work”).

Of course, there are other explanations for these errors, such as the interviewer keying in on certain words and not others (“disabled” but not “ill” for example) or interviewers choosing “other” when they were unsure if an answer matched a listed response.

Figure 2
What Is the Main Reason You Are Not Working Now?

Single-Response Item
11 Listed Responses
Complexity Rating: Medium

What is the main reason you are not working now?

ILL OR DISABLED AND UNABLE TO WORK	1	PREGNANCY/CHILDBIRTH	8
RETIRED	2	ON LAYOFF (TEMPORARY OR INDEFINITE)	9
TAKING CARE OF HOME/FAMILY/CHILDREN	3	JOB ENDED	10
GOING TO SCHOOL	4	NEW JOB TO BEGIN WITHIN 30 DAYS	11
CANNOT FIND WORK	5	OTHER (SPECIFY)	12
SUITABLE JOB NOT AVAILABLE	6	DON’T KNOW	d
NOT INTERESTED IN WORKING	7	REFUSED	r

Error Types and Rates	**Percentage (of 261 total files)***
No error	88.5%
Error rate	11.5%
Recorded a listed response as an "other" response	5.0%
Selected wrong code	2.7%
Recorded an "other" response as a listed response	2.3%
Entered “other specify” text that does not match the response	1.5%

*Percentages may total more than 100 because multiple codes were allowed.

4.2.3 When (CHILD) Misbehaves, What Do You Do?

The third question, a multiple-response item with 14 listed responses of high complexity, asked “When your child misbehaves, what do you do?” (see Figure 3). The response list for this item was complex, having a high number of words in the response labels, and responses with broad meaning (such as, “ground child/don’t let him/her go out or go out to play”). To maintain the momentum of the interview, the interviewer needed to be highly familiar with the list and the meaning of the codes. Not surprisingly, this question had a higher error rate (15.4%) than the previous two questions.

The most common error involved *selecting an insufficient number of codes*, found in 6.3% of files. This question typically elicited lengthy answers from respondents, which caused some interviewers to miss one or more appropriate codes. For example:

- Respondent said, “I tell her it’s wrong, of course. I yell…I talk to her…I take away privileges.” The interviewer correctly coded “3” (“talk to child”) and “12” (“take away privileges”) but failed to record “4” (“scold or yell at child”).

Figure 3
When (CHILD) Misbehaves, What Do You Do?

Multiple-Response Item
14 Listed Categories
Complexity Rating: High

Sometimes children misbehave. When (CHILD) misbehaves, what do you do? In other words, how do you punish or discipline (CHILD)?

GROUND CHILD/ DON'T LET HIM/ HER GO OUT/ GO OUT TO PLAY	1	TELL OTHER PARENT	10
SPANK CHILD	2	TAKE AWAY ALLOWANCE	11
TALK WITH CHILD	3	TAKE AWAY TV OR OTHER PRIVILEGES	12
SCOLD OR YELL AT CHILD	4	GIVE CHILD A "TIME OUT"	13
GIVE HIM/HER HOUSEHOLD CHORES	5	HOLD CHILD UNTIL (HE/SHE) IS CALM	14
IGNORE IT	6	OTHER (SPECIFY)	15
PUT CHILD IN ROOM/ SEND CHILD TO ROOM FOR LESS THAN 1 HOUR	7	DON'T KNOW	d
PUT CHILD IN ROOM/ SEND CHILD TO ROOM FOR MORE THAN 1 HOUR	8	REFUSED	r
MAKE CHILD GO TO BED	9		

Error Types and Rates	**Percentage (of 286 total files)**
No error	84.6%
Error rate	15.4%
Selected insufficient codes	6.3%
Selected wrong code	3.9%
Recorded a listed response as an "other" response	3.9%
Entered "other specify" text that does not match the response	0.7%
Failed to capture the entirety of the "other specify" response	0.4%
Recorded an "other" response as a listed response	0.4%

4.2.4 Why Would You Have (Little Interest/No Interest) in Attending These Classes?

The fourth question asked, "Why would you have (little interest/no interest) in attending these classes [that would help you have a healthy relationship with a spouse or partner?"]. This question was a multiple-response item with 10 listed responses (see Figure 4). We judged the response list to be the most complex because it contained response titles that were wordy, broad in meaning ("other family responsibilities"), and not well differentiated ("don't need or want services" and "just not interested"). In addition, the question was created for this survey, and the response list was untested. The error rate for this question was quite high, 23.8%.

The most common errors were *selected the wrong code*, and *recorded an "other" response as a listed response.* Each of these errors was found in 6.4% of all files. Examples of these errors included the following:

- Respondent said, "…I don't like nobody knowing my business," which should have been coded as "6." However, the interviewer coded the response as "8" ("don't need/want services/relationship fine").
- Respondent said, "I'm not in a relationship right now, so I don't need them," which should have been coded as "8." However, the interviewer coded the response as "10" ("just not interested").
- Respondent said, "I believe people can work it out sometimes talking to each rather than going to a class," which should have been entered as an "other specify" response. However, the interviewer coded the response as "10" ("just not interested").

Figure 4
Why Would You Have (Little Interest/No Interest) in Attending These Classes?

Multiple-Response Item
10 Listed Responses
Complexity Rating: High

Why would you (HAVE LITTLE INTEREST/NO INTEREST) in attending these classes?
(Free classes or workshops in your neighborhood that would help you [strengthen your marriage with (SPOUSE)/improve your relationship with (PARTNER)])

NEED TO WATCH CHILDREN	1	DON'T NEED/WANT SERVICES/RELATIONSHIP FINE	8
OTHER FAMILY RESPONSIBILITIES	2	CURRENTLY RECEIVING MARRIAGE/ RELATIONSHIP SERVICES FROM ANOTHER SOURCE	9
NO TIME	3	JUST NOT INTERESTED	10
LACK TRANSPORTATION	4	OTHER (SPECIFY)	11
WORK INTERFERES	5	DON'T KNOW	d
PRIVACY CONCERNS/MY OWN BUSINESS	6	REFUSED	r
SPOUSE NOT INTERESTED/ WOULD OBJECT	7		

Error Types and Rates	**Percentage (of 282 total files)***
No error	76.2%
Error rate	23.8%
Selected wrong code	6.4%
Recorded an "other" response as a listed response	6.4%
Selected insufficient codes	3.9%
Entered "other specify" text that does not match response	3.6%
Recorded a listed response as an "other" response	2.8%
Did not capture entirety of "other specify" response	1.8%
Entered "other specify" text that was unintelligible/ not usable	0.4%

*Percentages may total more than 100 because multiple codes were allowed.

5. Conclusions

In this study, we used CARI to study a particular type of interviewer error, namely, data capture error when recording responses to open-ended questions with precoded response lists. Overall error rates depended on the structure and complexity of the question and response list. We found that the question *How is this person related to you?*—a cognitively simple, single-response question—had the lowest overall error rate. In contrast, the question *Why are you not interested in taking these classes?*—a multiple-response question with a conceptually difficult response list, had an error rate that was almost four times as high. The length of the response list did not seem to make a difference; rather it was the labels of the codes themselves, the interviewers' lack of familiarity with the response lists, and the interviewers' inadequate understanding of the meaning of the responses that led to the errors. These findings point to the need to train interviewers well in the meaning and use of response lists and to include examples from actual field experience to maximize relevance.

Across all questions, *selected the wrong code* was the first or second most common error. Based on our qualitative assessment of the audio recordings, these errors had to do with both keying error and with interviewers misinterpreting the codes or not taking the time to find the best match. Indeed, we heard evidence of interviewers "satisficing," that is, selecting a code that is appropriate, but not necessarily optimal. Usually the need to maintain the flow of the interview and the respondent's attention was behind the decision to select the first code that seemed to match, rather than the best code(s). Further, we observed that rather than slow down the interview to consider the codes, interviewers would opt to record the response in "other" as a way to maintain the interview pace.

Of course, interviewers are not solely at fault for the errors we observed. The response descriptions were sometimes overlapping and ambiguous, making it difficult for even highly skilled interviewers to differentiate among them. We found, for example, that two responses to the question *Why are you not working now?* were particularly confusing: "cannot find work" and "suitable job not available." If the respondent answered, "I look every day but I can't find anything interesting," then reasonable interviewers might disagree on how to code that response, leading to inconsistent coding across interviewers. This finding points to the need to create response labels that are clear and descriptive, rather than rely on interviewers to remember code definitions from the training manual.

Only one of the four items (Why are you not interested in taking the classes?) was newly created for this survey, whereas the other three items were previously tested and used in other surveys. This question had the highest coding error rate. Although the response list was rated highest in complexity, the fact that it was created without knowledge of how respondents would answer might mean the listed responses were inappropriate, too broad (leading to errors of inclusion), or not mutually exclusive. The untested design may well have contributed to errors of interpretation or classification that inflated the error rate to this question.

Although based on evidence from only four questions, results from this study led us to the following conclusions:

- Multiple response questions have higher coding error rates. This finding is partly due to there being more opportunities to make errors by selecting insufficient codes to fully capture a respondent's answer.
- Questions that are more cognitively complex for respondents are more prone to data capture error because respondents may not answer in a clear and logical way, making more error-prone the interviewers' task of parsing and matching the response.
- There is some evidence of "satisficing" by interviewers, in which they select a code that is appropriate--but not the most appropriate.
- Questions used for the first time, where the response list is contrived without prior experience, may be more prone to data capture error.
- Entering inaccurate or incomplete "other specify" responses and recording responses that are unintelligible/not codeable are not common errors.

We conclude that some questions with high error rates might be better asked as true open-ended questions, in which the interviewer records the respondent's answer verbatim. The verbatim answers are then coded by trained coders with full access to written code definitions and free from the pressure of the interview environment. Moreover, performing a post-survey review of "other specify" responses can reduce overall coding error by changing "other specify" answers to listed responses when appropriate. For the questions in this study, we undertook this review and reduced error rates by 1 to 5 percentage points, depending on the question.

In sum, CARI is a useful tool for identifying problem questions, deciding which questions are better asked as true open-ended questions, estimating interviewer error rates of different kinds, and abstracting actual examples of interviewer behavior to improve interviewer training in the future. Although expensive, it is possible to correct many of the data capture errors we observed if a project has the time and resources to conduct a complete review of the audio files. Performed on a sample basis, this quality control measure would be an affordable and useful step for new data collections, as well as existing data collections interested in learning more about interviewer error.

References

Collins, W. A. (1970). "Interviewers Verbal Idiosyncrasies as a Source of Bias." Public Opinion Quarterly, Vol. 34, No.3., pp. 416–422.

Dielman, L. and Couper, M. (1995). "Data Quality in CAPI Surveys: Keying Errors." Journal of Official Statistics, Vol. 11, pp. 141–146.

Fowler, F. J. and Mangione, T. W. (1985). "The Value of Interviewer Training and Supervision." Final Report to the National Center for Health Services Research, Grant #3-R18-HS04189.

Groves, R. (1989). Survey Errors and Survey Costs. New York: John Wiley and Sons, Inc.

Groves, R. (1987). "Research on Survey Data Quality." The Public Opinion Quarterly, Vol. 51, Part 2: Supplement: 50th Anniversary Issue, pp. S156-S172.

Kennedy, J. M., Lengacher, J. E., and Demerath, L. (1990). "Interviewer Entry Error in CATI Interviews." Paper presented at the International Conference on Measurement Errors in Surveys, November, Tucson, AZ.

Mangione, T. M., Fowler, F. J., and Louis, T. A. (1989). "Question Characteristics and Interviewer Effects." Journal of Official Statistics, Vol. 8, pp. 293–307.

Rustemeyer, A. (1977). "Measuring Interviewer Performance in Mock Interviews." Proceedings of the Social Statistics Section, American Statistical Association, pp. 341–346.

Thissen, M. R., Sattaluri, S., McFarlane, E. S., and Biemer, P. P. (2007). "Evolution of Audio Recording in Field Surveys." Journal of Survey Practice.