Article

Computer Audio-Recorded Interviewing (CARI): A Tool for Monitoring Field Interviewers and Improving Field Data Collection

by M. Rita Thissen, Carl Fisher, Lillie Barber
    and Sridevi Sattaluri

2009

Statistics Canada   Statistique Canada

Canada

Computer Audio-Recorded Interviewing (CARI): A Tool for Monitoring Field Interviewers and Improving Field Data Collection

M. Rita Thissen, Carl Fisher, Lillie Barber and Sridevi Sattaluri [1]

Abstract

Survey managers are still discovering the usefulness of digital audio recording for monitoring and managing field staff. Its value so far has been for confirming the authenticity of interviews, detecting curbstoning, offering a concrete basis for feedback on interviewing performance and giving data collection managers an intimate view of in-person interviews. In addition, computer audio-recorded interviewing (CARI) can improve other aspects of survey data quality, offering corroboration or correction of response coding by field staff. Audio recordings may replace or supplement in-field verbatim transcription of free responses, and speech-to-text technology might make this technique more efficient in the future.

Key Words: CARI, Audio recording, Quality assurance, Data collection, Transcription.

1. Monitoring data collection

The primary goal of monitoring data collection is to ensure the quality of the resulting data, taking into account anything that may contribute to total error. A second but still important goal is to manage the performance and integrity of field staff with respect to job requirements. The second goal interacts with the first, since a data collector who exhibits unprofessional behavior or cheats on the business aspects of the job may also return biased or falsified data.

Three general aspects of data collection can influence the quality of the resulting dataset:
- Conformance to protocol by interviewers
- Authenticity of data capture by interviewers
- Validity of data with respect to the intent of the survey.

CARI techniques can help manage and control error from each of these sources

Recognizing that data problems can result from field performance, most survey organizations follow standard practices to monitor interviewers' work. Through the years, many approaches have become common. Table 1-1 summarizes some of these and their advantages and disadvantages. Now there are additional tools and techniques for ensuring quality though some are not yet ready for wide-scale production usage, such as video monitoring.

The discussion of CARI below describes the state of the art for monitoring quality through audio recording, standing on the shoulders of existing methodological knowledge and leveraging advances in computerization.

---

[1] M. Rita Thissen (rthissen@rti.org), Carl Fisher (carlf@rti.org), Lillie Barber (lbb@rti.org) and Sridevi Sattaluri (ssattaluri@rti.org), RTI International, PO Box 21294, Research Triangle Park, NC 27709, USA

Table 1-1
Classic techniques of monitoring field data collection

Technique	Advantages	Disadvantages
In-person observation	Provides excellent detail	May affect interviewer's or respondent's conduct May be costly
Post-interview discussions with interviewers	Easy Inexpensive	Provides only the interviewer's viewpoint
Verification contact, by telephone or in-person re-interview	Obtains respondent's point of view	Adds burden on the respondent May be costly
Review of response data and timers	Effective for monitoring data quality	Limited evidence about performance
Tape recording during interview	Informative	Obtrusive Awkward if tapes must be stopped and started Extra equipment costs Logistical effort of shipping tapes

2. CARI overview

CARI is a set of software and methodological practices and tools which collect audio recordings of the interviewers' and respondents' vocal exchanges during an interview for later review by quality monitors or others. Its main use to date has been for quality control of field work, and the same technology offers other opportunities described in later sections of this paper.

Digital audio recording for field interviews was explored at RTI in 1998 and went into production use in 1999 as a questionnaire-controlled recording tool. The first implementation relied on the CASES (http://cases.berkeley.edu:7504/index.html) questionnaire language, C software and a SoundBlaster audio card. At that time, laptop computers were used for a much smaller fraction of field surveys than now, because hardware was expensive and heavy, disk capacity limited and processing power low. Still, CARI was demonstrated as a viable production technique on the National Survey of Child and Adolescent Well-Being (Biemer et al, 2000), which has continued to use it each year since then in subsequent rounds of the longitudinal survey. Other field surveys conducted by RTI have followed suit and continue to employ the approach.

3. CARI prevalence

Numerous organizations have presented research at conferences that have mentioned use of digital audio recording for surveys. The authors are aware of CARI for field work during or prior to 2008 at the following organizations (and there may be more):
- RTI International, USA
- U. Michigan, USA
- Battelle Institute, USA
- National Centre for Social Research, UK
- National Opinion Research Center, USA
- US Census Bureau (field tests only)
- Statistics Canada
- Statistics New Zealand
- Westat, USA

At Statistics New Zealand (Keefe 2008) and elsewhere, CARI is one component of a comprehensive quality plan that is being standardized for all field surveys. The Australian Bureau of Statistics is watching their efforts to determine whether it should adopt CARI technology also.

4. Quality findings through CARI

As mentioned above, survey quality concerns include conformance to protocol, the accuracy of data capture, the validity of responses, and job performance by interviewing staff. CARI technology can support quality control for each of these aspects of survey quality management.

How can survey managers confirm that interviewers are adhering to their training? By capturing all sound within approximately 5 to 8 feet in a CARI file, a monitor can listen to the exchange that took place during the interview, including the exact phrasing used by the field worker and the exact answer given by the respondent. At RTI, survey teams define problem codes such as the ones listed in Table 4-1 from the Survey of Community and Family Life (Thissen et al, 2007) to track the severity of deviation from planned delivery. As shown here, the vast majority of cases do not show problems, and of the problems which are detected, many are minor.

Table 4-1
Conformance to protocol, based on evaluation of 5600 cases

Count of Cases	Percent	Problem Description
13	0.2	Authenticity Questionable**
217	3.9	Reading – Minor Deviation
72	1.3	Reading – Major Deviation
86	1.5	Inappropriate Probing
79	1.4	Feedback not Neutral

** Mostly found to be all right, after additional investigation

On occasion, the recorded audio file may point to the possibility of data falsification, as shown in the first entry in Table 4-1. Several characteristics raise suspicion of curbstoning:
- Absence of a respondent's voice in the recording when other sound is present
- Inconsistent responses
- Omission of items or sections from the audio recording

That suspicion is grounds for a more thorough review of the data, including other CARI files from the same interviewer. Typically, the review indicates turns up an explanation which may be poor practice but is not outright falsification.

A high CARI refusal rate may be an indicator of curbstoning. If an interviewer is falsifying data by not conducting the interviews, he or she may hope to avoid detection by marking the CARI consent form with a refusal. CARI refusal rates are not conclusive evidence of falsification. Some regions of the country and some respondent populations may be more concerned with privacy than others. A high refusal rate may also indicate improper delivery of the consent question, such as following the informed consent with a statement, "You know you don't have to agree to it." Alternate methods are needed for confirmation of the authenticity of cases where CARI was refused, such as a sample for telephone verification calls or re-interview.

With a modicum of additional effort, CARI monitors can simultaneously listen for authenticity and also evaluate interviewer performance in order to provide positive or corrective feedback. This is an especially valuable tool for following up after training of new interviewers, as it can identify aspects of training which might need reinforcement. Rates of serious performance problems are low, such as negative feedback or inappropriate probing (noted in Table 4-1) and less than 1% of cases showing evidence of unprofessional behavior (Thissen et al, 2007). Positive feedback may be provided for many actions, including

- Precise adherence to protocol
- Adept handling of difficult situations
- Consistency, honesty, and professional behavior
- Appropriate paraphrasing

Interviewing staff understandably are pleased to receive positive feedback, especially for times when they have had to overcome resistance or obstacles to conducting the interview.

Aside from rating field performance and authenticity, CARI can provide estimates of data entry accuracy and the effect of question wording on the ability of interviewers to correctly capture response data. A recent study of complex multiple choice questions showed that some type of data entry error occurred on 14.5% of the studied questions, as shown in Table 4-2 (Mitchell et al, 2008).

Table 4-2
Correctness of data capture for multiple choice plus "Other, Specify"

Description	Percent (1,083 files)
No error	85.5%
Error total*	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%

*Individual error categories total more than 14.5% because multiple codes were allowed.

Some of the errors can be seen as benign, such as recording an answer in the "other, specify" field which properly belonged in one of the categorical choices. Such an error causes no loss of data, because post-data collection review allows corrective recoding.

This error rate may seem high, but it does not apply to simple questions with a few response options. The items under study had 12 to 19 response options each, including a field for recording open-ended text in case the respondent's answer did not fit one of the predefined categories. The details of this study indicated a trend of higher error rates associated with longer response lists and categories which might be subject to interpretation.

5. Questions and concerns

No methodology is without drawbacks. Some of the issues which have been raised about audio recording are easily addressed while others are more problematic. Legal issues such as the format and timing of informed consent must be decided on the basis of local and national regulations. Similarly, laws vary as to the admissibility of recordings in termination procedures. RTI has chosen to act conservatively on these issues, obtaining written consent and not relying on audio files as disciplinary evidence. In the event of falsification or other serious accusation, additional evidence was found to support the complaints, and CARI files were not used directly.

Is there an effect on response rate? Some sample members may be reluctant to allow recording, especially in surveys of a sensitive nature. RTI's approach is to obtain consent for the interview first and separately request approval for recording. Respondents can agree to participate in the survey without having to agree to CARI. We also provide an option that those who consent to CARI may change their minds at any point in the interview, at which time the interviewer can signal the instrument that permission has been retracted. Over many years and many thousands of interviews, CARI acceptance rates at RTI have been upwards of 80%, including surveys of very sensitive topics.

How much recording is enough? CARI serves different purposes, as discussed above. The number of CARI files, the duration of each recording and the selection of which items to record depend on the purposes for which the audio recordings will be used.

The minimum number and duration was estimated in a study conducted by RTI with funding from the US Census Bureau, in which researchers determined that reviewing three 30-second files was enough for pairs of independent raters to converge on a rating of authenticity (Thissen, Sattaluri, McFarlane and Biemer, 2007). At the other extreme, some researchers record entire interview sessions for later review; this approach is particularly valuable with semi-structured interviewing techniques. With compression, three minutes of voice recording can be

transmitted over dial-up lines in approximately three minutes of connection time, a concern which is especially important in areas that lack high speed internet.

RTI's practice has been to record a minimum of three segments of one minute each, chosen from the beginning, middle and end of the interview, and selecting items that will be asked of all respondents, not bypassed by skip logic. Generally, survey teams elect to record additional questions, such as key points of information such as gateway questions or confirmation of the receipt of incentives, questions with "other, specify" options, or questions in which the content is of critical importance to subject matter experts and analysts.

How long should the files be stored? Cost, informed consent wording and privacy issues determine the answer to this question.

File storage prices have dropped precipitously over the past decade, allowing long term storage of large data or audio files for low cost. In CARI's early implementation, storage costs mandated quick removal of audio files after review. However, there can be a benefit of long term storage, such as when an interviewer becomes suspected of serious falsification and quality specialists would like to review a complete interviewing history to determine when curbstoning started.

The exact wording of the informed consent information may impact an organization's authority to retain and use CARI files. Clearly, if the informed consent wording dictates file removal within a given timeframe, that limit must be followed. Similarly, if the wording specifically identifies the recording for evaluating interviewer performance, files cannot be retained for methodological studies or examination of questionnaire performance. A simple consent phrase can provide the greatest flexibility such as, "This interview may be recorded for quality purposes. Do you agree?"

What is the effect on the morale of field staff when CARI is introduced? Change management is always problematic, and the introduction of new technology is no exception. A pretest conducted by the US Census Bureau indicated that some interviewers might resent the change, especially those with a long history of good performance. That same study found that the reaction varied greatly from one regional office to another, possibly depending on the attitudes of the senior managers of the office (Arseneaux and Thorpe, 2007). A debriefing of interviewers in 1999 at RTI, when the technology was introduced, indicated that 82% were either neutral or positive about the use of audio recording (Biemer et al, 2001)

6.   Future directions

The best technological advances are those which offer immediate value and also open new avenues for future benefit. Use of audio recording in field interviews provides an instant return, and it appears that many additional methodological directions may be supported by its introduction.

Westat and the Blaise Users Group have organized a working group to explore the addition of audio recording capabilities to Blaise software (Statistics Netherlands, www.cbs.nl/blaise/). This group is expected to begin discussions in 2009. Possible options include enhancement to the Blaise audit DLL (dynamic link library) for communication with external sound recording or making available code examples in the Blaise help files.

Many possibilities exist for using CARI for methodological studies. Some of these were discussed earlier for evaluating the effectiveness of questions, for examining translation effects, as a production-phase option for the equivalent of laboratory focus groups and for assessing interviewer recording accuracy. Surveys have largely been restricted to multiple-choice responses since the introduction of laptops, but use of audio recording might re-open the option of capturing conversational responses for later behavior coding or transcription at a central location. CARI is only a tool, but like other technological changes, it presents opportunities that were not possible previously.

It is likely that recording applications similar to CARI will come into use on smaller mobile devices such as handhelds and smart phones which have recording capabilities. Instrument languages, file storage, case management and transmission obstacles affect the use of these devices for surveys in general and the ability to capture and manage audio data.

7. Speech recognition and transcription

One promising area for exploration is that of automated speech transcription, commonly called "speech to text" software. Several well-known transcription packages are available for desktop users, including IBM's Via Voice, Dragon Naturally Speaking and Microsoft's speech recognizer delivered as part of Word data processing software. These products are not so valuable for general use on surveys, for several reasons:

- Requirements for hardware such as specific headset microphones
- Requirement for training the software to recognize each voice before automated use
- Requirement for training the speakers to provide very uniform inflection
- Intolerance of background noise
- Intolerance of multiple voices in a single session
- Intolerance of accents, dialects and specialized vocabulary

Few field surveys can meet the restrictions without hindering the flow of the survey.

However, Carnegie Mellon University began a software project called Sphinx (The Sphinx Project, cmusphinx.org/) which has grown into an international open-source effort to make audio operations amenable to real-life situations. Its basic elements include an acoustic model that converts sound to phonemes, a grammar of known words and phrases and their phonetic equivalents, and a language model that associates probabilities with those words and phrases.

Table 7-1
Mock CARI files used as standards for testing software

Characteristic	Sample Population
Sex	11 male, 21 female
Age	6 aged < 40, 15 aged 40-49, 11 aged 50+
Pitch	5 low, 26 moderate, 1 high
USA	23 mostly or always lived in USA, 9 did not
English	24 spoke English at home age 1-14, 8 did not
Loudness	4 quiet, 25 medium, 3 loud
Speed	5 slow, 25 moderate, 2 fast
Room Noise	22 quiet, 9 moderate, 1 loud

RTI has begun investigating the use of Sphinx software to see if it could automatically transcribe CARI files. We started prepared a library of audio files for which we knew voice characteristics, collected from mock interviews of volunteer staff. The questionnaire asked interviewers and respondents to rate their own voices with respect to attributes shown in Table 7-1.

Our first attempt was highly frustrating. Out of 524 transcriptions, only 5 were accurate enough to be considered good to excellent, and only 12 were rated at medium quality or higher (Thissen et al, 2008). However, postings on the open-source forum encouraged us to proceed, and our most recent configuration (Trial 5) resulted in 40% of the transcriptions good enough to be rated as perfect or nearly perfect and an additional 30% rated as fairly good, as shown in Table 7-2.

Table 7-2
Preliminary results from using Sphinx for transcription

Count of Cases	Percent	Outcome
193	40%	Perfect or nearly perfect
149	30%	Medium (a few words missed, added or incorrect)
144	30%	Poor (many words missed, added or incorrect)

Some of the transcriptions in the "perfect" category were as much as 50 words long. For example, the item shown in Table 7-3 from the Blaise specification document was transcribed perfectly for more than one speaker, in spite of its length. Some of the "medium" files showed problems transcribing non-language sounds (sneezes, "uh",

"hmmm"). Such imperfections may be acceptable for automated use, since the non-speech noise is irrelevant to our purpose.

Table 7-3
A question which was transcribed perfectly for some speakers

Blaise Specification Text
"We are at the end of the interview. We have been asked to comment on any unexpected sounds within the interview session, such as interruptions, sudden loud noises, technical problems or anything else which might affect the audio recordings. I would say   @/@/(INTERVIEWER COMMENT, SUCH AS ""There were no problems that I can recall."")"

However, the same item was not always transcribed accurately. For example, in one mock interview the speaker had an Australian accent, did not read the words exactly, overlapped with the respondent's voice, and trailed away to silence at the end. See Table 7-4. It would be appropriate for an automated process to flag this recording for review, and a monitor would find performance faults as well as transcription faults.

Table 7-4
Poor match between specification and transcription

Spoken
Ah, we are at the end of this interview. We have been asked to comment on any unexpected sounds within the interview session, such as interruptions, sudden loud noises, technical problems or anything else which might affect the audio recordings. I would say, ah, there were no problems that I can recall (sniff). So…<background speaker>There were no problems. So I <unintelligible>
Transcribed
UNITED END OF THE INTERVIEW WE HAVE BEEN ALL TO COMMENT ON ANY UNEXPECTED SOUNDS SPEAKING BEING TO BE SESSION SUCH AS INTERRUPTIONS SUDDEN LOUD NOISES TECHNICAL PROBLEMS OR ANYTHING ELSE READ TO QUIET IF A THE AUDIO RECORDINGS I WOULD SAY ON THERE WERE NO PROBLEMS THAT I CAN RECALL AT A THAT I CAN AT THE END OF THE

8. Summary

In summary, digital audio recording is another helpful tool for survey research. It improves survey managers' abilities to evaluate the performance of field staff rapidly and inexpensively, and it also allows evaluation of the questions themselves. At the same time, use of CARI reduces burden on respondents who might otherwise be re-contacted.

Use of the technology has spread over the past several years and is likely to continue growing. Current field tests and system development are taking place at major survey organizations in several countries.

Current usage is largely tied to confirming authenticity of interviewers and providing performance review for field staff. Future possibilities include greater use for capturing open-ended responses, for assessing question wording and other aspects of questionnaire design and evaluation.

Enhancement and refinement of techniques can be expected, and other technologies may build on the CARI foundation in the coming years, such as the prospect of automated transcription of the collected recordings.

References

Biemer, P.P., Hergert, D., Morton J. and Willis W.G. (2000). The Feasibility of Monitoring Field Interview Performance Using Computer Audio Recorded Interviewing (CARI), *Proceedings of the American Statistical Association's Section on Survey Research Methods*, 1068-1073

Biemer, P.P., Mierzwa, F. and Thissen, M.R. (2002). Comparison of the Steady-State Costs for CARI vs. Reinterview Approaches for Interview Verification, RTI Technical Report to the US Census Bureau

Egan, M. (2007). Statistics Canada, personal communication.

Fowler, F. J. (2001). *Survey Research Methods*, SAGE Publications Inc., Thousand Oaks, CA USA, p.6.

Herget, D., Biemer, P.P., Morton, J. and Sand K. (2001). Computer Audio Recorded Interviewing (CARI): Additional Feasibility Efforts of Monitoring Field Interviewer Performance, *Federal Conference on Survey Methodology (FCSM)*, Washington, DC USA

Keefe, S. (2008). Statistics New Zealand, personal communication.

Mitchell, S.B., Fahrney, K.M., Strobl, M.M., Thissen, M.R., Nguyen, M.T., Bibb, B.S. and Stephenson, W.I. (2008). Using Computer Audio-Recorded Interviewing to Assess Interviewer Coding Error, *63rd Conference of the American Association for Public Opinion Research (AAPOR)*. New Orleans, LA USA

Thissen, M.R., Sattaluri, S., McFarlane, E.S. and Biemer P.P. (2007). Evolution of Audio Recording in Field Surveys, *62nd Conference of the American Association for Public Opinion Research (AAPOR)*, Anaheim, CA USA

Thissen, M.R, Sattaluri, S. and Fisher, C. (2008). Will Speech-to-Text Software Work on Audio Recordings from Field Data Collection?, *International Field Directors and Technologies Conference*, New Orleans, LA USA