Ensuring Data Quality: Monitoring Accuracy and Consistency Among Telephone Interview Monitors

Joseph Baker, Claudia Gentile, Jason Markesich, Shawn Marsh, and Rebecca Weiner

Joseph Baker, Mathematica Policy Research, 600 Alexander Park, Princeton, NJ 08540
Claudia Gentile, Mathematica Policy Research, 600 Alexander Park, Princeton, NJ 08540
Jason Markesich, Mathematica Policy Research, 600 Alexander Park, Princeton, NJ 08540
Shawn Marsh, Mathematica Policy Research, 600 Alexander Park, Princeton, NJ 08540
Rebecca Weiner, Mathematica Policy Research, 600 Alexander Park, Princeton, NJ 08540

Abstract

Although interviewer monitoring is central to survey quality assurance, little research has examined the accuracy and consistency of monitors' behaviors. In 2010, Mathematica Policy Research conducted an exploratory study that found a high degree of accuracy and consistency among monitors' overall interview ratings, but a limited use of the rating scale (Baker et al. 2010). We conducted a follow-up study to further examine the following questions: (1) How accurate and consistent are monitors' overall ratings of interviews? and (2) How accurate and consistent are monitors in identifying nonstandardized interviewer behaviors?

We examined the behaviors of two groups of monitors: three senior monitor supervisors and 12 active monitors who each evaluated 20 digitally recorded interviews from six projects. Monitors used our five-point Interviewer Rating Scale to assign an overall rating, and a behavioral coding system to highlight both positive and nonstandardized interviewer behaviors. To measure accuracy and consistency, we compared monitors' ratings within and across the two groups and compared their behavior codes at the question level. To examine the extent of monitor variation within each monitor over time, we asked monitors to reevaluate three interviews from the 2010 study.

Analyses of monitors' overall ratings of interviews revealed that both the monitor supervisors and active monitors were very consistent in their ratings, and the active monitors were consistent with the monitor supervisors. Also, monitors were consistent with themselves over time. However, there was little variation in their ratings; monitors mostly assigned ratings of "2" (does not meet expectations) and "3" (meets expectations). Overall, both groups used the nonstandardized behavioral codes in ways consistent with their overall ratings. Monitors tended to focus on different parts of the interview when assigning their nonstandardized behavioral codes, but when they did assign codes to the same part of an interview, these codes were consistent.

Keywords: monitoring, consistency, training, telephone interviewers, interviewer behavior, quality control, data quality

1. Introduction

For survey research organizations, monitoring telephone interviewers to ensure that they follow standardized interviewing procedures is a standard part of data quality assurance. At Mathematica Policy Research, we monitor observations to identify problems with survey questions, recommend interviewing techniques, conduct methodological investigations of questionnaire designs, and retrain interviewers whose performance does not meet expectations. Because monitoring is a critical quality assurance tool, we are interested in understanding and assessing the behavior of monitors—specifically, their ability to provide effective and consistent feedback on interviewer performance. Mathematica implements several best practices to promote monitoring consistency:

- Observing at least 10 percent of each interviewer's work using a standardized monitoring form and rating scale
- Dedicating staff with previous interviewing experience to monitoring activities
- Providing comprehensive training for monitors
- Providing immediate feedback to interviewing staff on aspects or techniques that were performed well during the interview and areas that need improvement
- Producing statistics on the average evaluation scores, interviewing errors, and percentage of hours monitored by interviewer and project

Despite the implementation of these best practices, anecdotal evidence from interviewers and monitors suggests that monitors are not always consistent in how they evaluate interviews. For example, interviewers have noted that different monitors tend to focus on different nonstandardized interviewing behaviors (such as changing the wording of questions, data entry and coding errors, reading questions too fast, and probing errors) when evaluating interviews and providing feedback. Interviewers also note that some monitors are more stringent than others in terms of the criteria they use to rate an interview. For example, some monitors seem reluctant to rate an interview as above average or excellent. These types of variations in monitoring behaviors could have an impact on data quality, the reliability of interviewer performance ratings, and staff morale and retention. More specifically:

- If monitors emphasize certain interviewing behaviors at the exclusion of others, or treat interviewing errors differently, the quality of telephone interviews might be compromised. This is especially problematic if one monitor glosses over behaviors deemed unacceptable by another.
- If the monitors provide conflicting feedback, or focus more on negative behaviors than on the positive aspects of the interview, telephone interviewers might become discouraged or resign.

Research on understanding monitor behavior or effects is not extensive. Most studies have focused on describing monitoring processes or methods, such as the key elements of an effective monitoring system (Cannell and Oksenberg 1988; Fowler and Mangione 1990; Lavrakas 2010), or how organizations monitor the quality of their work (Burks et al. 2006; Steve et al. 2008). Tarnai (2007) discussed the advantages and disadvantages of monitoring both complete and partial interviews and examined interviewers' reactions to the monitoring process. Other studies explained the development and use of standardized monitoring forms and/or scoring procedures to measure the performance of telephone interviewers (Sudman 1967; Couper et al. 1992; Mudryk et al. 1996; Currivan et al. 2006;

Durand 2005; Steve et al. 2008). Mathematica previously conducted an exploratory study of monitors' consistency and accuracy, indicating a need for a more in-depth look at the monitoring process (Baker et al. 2010).

Thus, little is known from research about the factors affecting monitors' judgments and behavior that could inform the improvement of interview quality control procedures. To explore monitor behavior, Mathematica designed a monitoring consistency exercise that addressed the following questions:

- How accurate and consistent are monitors' overall ratings of interviews?
- How accurate and consistent are monitors in identifying nonstandardized interviewer behaviors?

We asked two groups of monitors (gold standard and active monitors) to evaluate eight digitally recorded interviews from three telephone surveys. This paper presents the key findings of this study. Section 2 provides background on our monitoring system, form, and rating scale; Section 3 describes the methods used in this research; Section 4 presents the results of the study; and Section 5 discusses the implications of the results.

2. Monitoring System, Form, and Rating Scale

Central to the quality assurance process is a monitoring system that enables monitors to listen unobtrusively to telephone interviews and view an interviewer's computer screen while an interview is in progress. In addition, digital recordings of interviews provide monitors with a tool for monitoring at any time. Mathematica monitors regularly review digital recordings with interviewers to discuss aspects of their interviews that need improvement. We inform interviewers that we will monitor them, but they do not know when observations will take place; they can be monitored randomly or at the discretion of project staff. The monitors evaluate interviews using an electronic monitoring form composed of the following sections:

- **Session information.** We collect the following information: monitor's name; interviewer's name; project; date; start and end time of the monitoring session; selection type (probability selection, supervisor request, interviewer is new to project); and whether the monitor evaluated a complete interview, a partial interview, or only an introduction.
- **Behavioral coding system.** A summary of the non-standardized and positive interviewing behaviors observed during the course of the interview. When an interviewer makes an error or does something very well, the monitor enters the question number, behavioral code, and any relevant comments. Monitors select from 17 behavioral codes across five categories: (1) errors in reading questions, (2) probing errors, (3) feedback errors, (4) coding/data entry errors, and (5) positive comments.
- **General voice and rapport.** The monitor evaluates the interviewer's volume, pace, clarity, tone, and rapport, assigning a code of standard (voice characteristic was appropriate) or nonstandard (voice characteristic was deficient).
- **Administration of pre- and post-questionnaire tasks.** The monitor notes whether or not the interviewer accurately introduced the study properly and recorded the callback date/time, call disposition, and interviewer notes.

- **Comments on overall performance.** The monitor briefly summarizes aspects or techniques performed well during the interview, aspects or techniques that need improvement, and a plan of action for future interviews.

After completing the monitoring form, the monitors assign an overall rating for the session, using the five-point scale presented below in Figure 1.

Rating	Description	Definition
1	Unacceptable	Needs immediate supervisor attention, possible grounds for termination. Many errors of a serious nature (i.e., falsifying data, abusive or unprofessional feedback, skipping questions).
2	Does Not Meet Expectations	Interviewer needs further monitoring. Several significant errors (i.e., major wording changes, leading probes, biasing responses, introducing the study in an inappropriate/inaccurate manner, coding errors).
3	Meets Expectations	Straightforward interview with a typical respondent that meets standards. Very little probing, rereading or answering of respondent's questions needed. Very few or insignificant errors (i.e., minor probing or spelling errors or minor wording changes).
4	Very Good	Challenging interview involving a fair amount of probing, rereading of questions, or typing of open-ended/verbatim responses, all of which were done accurately. No errors or only a few insignificant errors.
5	Excellent	Very challenging interview requiring a great deal of probing or rereading of questions. The interviewer might have converted a hard-core refusal or kept a respondent with a physical or cognitive impairment on track during the interview. No errors or one minor error.

Figure 1: Interviewer Rating Scale

3. Methodology

In an effort to address our research questions and improve our quality assurance procedures, we conducted the monitoring consistency exercise during January, February, and March 2011. In this section, we describe the subjects and materials used to carry out the study, data collection procedures, and methods of analysis.

3.1 Subjects

Twelve active monitors and three supervisors were recruited for this study and asked to evaluate 20 digitally recorded telephone interviews conducted by a cross-section of interviewing staff. The 12 active monitors represented a range of experience, with 1 to 17 years of experience interviewing and monitoring. The three supervisors had 4 to 10 years of experience interviewing and monitoring and 3 to 5 years of experience as supervisors. They served as the gold standard; their ratings were the criteria used to judge the ratings of the active monitors.

When both the gold standard monitors and active monitors first became monitors, they received specialized training on Mathematica's monitoring procedures and systems. Their training included an in-depth introduction to the monitoring system, procedures for how

to apply monitoring standards consistently, and guidelines for providing constructive feedback to interviewers. During the final stage of training, experienced monitors closely supervised newly trained monitors. This process is designed to ensure that all monitors fully understand the monitoring systems and evaluation scale and provide feedback in an objective and constructive manner.

3.2 Selecting Interview Sessions to Evaluate

During the course of a given day, monitors evaluate interview sessions that vary by study content and respondent populations, interviewer skill level, interview length, and session type (complete interview and partial interview). Therefore, we selected a mix of digital recordings based on these characteristics. First, we identified projects that offered a range of topic areas and respondent populations. Of the projects that were in the midst of data collection at the time of our study, we selected digital recordings from (1) Building Strong Families (BSF) (parents interviewed about their relationship with their partners); (2) Evaluation of Individual Training Account Demonstration (ITA) (customers interviewed about training voucher programs); (3) The Early Head Start Family and Child Experiences Survey (BabyFACES) (parents interviewed about their children's experiences with the EHS program); (4) Community Tracking Study (CTS) (households interviewed about their health care); (5) National Beneficiary Study (NBS) (customers interviewed about Social Security disability benefits); and (6) National Mental Health Services Survey (N-MHSS) (survey of mental health treatment facilities).

To increase the likelihood that the interviews used in the study would vary in terms of quality, we then identified interviewers with different skill levels. For each project, we reviewed the monitoring reports, and classifying interviewers as either above average (those with average ratings above 3); average (those with average ratings of 3); and below average (those with average ratings below 3). We then randomly selected six above-average, seven average, and seven below-average interviewers from the pool of interviewers engaged in the six projects mentioned above.

Because monitors evaluate both complete and partial interviews, we included 15 complete interviews and five partial interviews, each of which contained an introduction. Lastly, we selected one digital recording from each of the twenty interviewers, taking into consideration the need to select a mix of complete and partial sessions. Table 1 provides a summary of the selected recordings.

Table 1: Interviews Selected for the Study

Number and Length of Interview	Type of Interviewer
15 Completes	
Two 10–20 minute interviews	Six Above Average
Eight 21–30 minute interviews	Five Average
Three 31–40 minute interviews	Four Below Average
Two 41–50 minutes interviews	
5 Partials	
Four 10–20 minute interviews	Two Average
One 31–40 minute interview	Three Below Average

To examine the extent of variation within each monitor, three of the interviews from our 2010 pilot study were re-rated by 9 monitors in the 2011 study, yielding 27 total observations. This allowed us to analyze within-monitor consistency in their overall ratings and non-standard behavior codes.

3.3 Data Collection

To carry out the monitoring consistency exercise, during a six-week period we scheduled individual monitoring sessions with each study group: the 12 active monitors and the three gold standard monitors. During the first meeting with each group, we informed the study participants that the purpose of the exercise was to gather data that would help us improve the monitoring form and process. We also informed the participants that they would monitor and evaluate the digitally recorded interviews independently of one another, and that they were not permitted to discuss how they rated the interviews with their colleagues. Both the active monitors and gold standard group monitored the digital recordings and summarized nonstandardized interviewing behaviors and positive aspects of the interviews in a monitoring database. They each evaluated the 20 digital recordings, yielding a total of 300 observations for the study.

4. Data Analysis

To address the question of how accurate and consistent the monitors are in their overall ratings of interviewers, we compared the ratings among the gold standard monitors, among the active monitors, and the overall agreement among all of the monitors. We assessed the accuracy of the active monitors by comparing their overall ratings to those of the gold standard group. To address the question of what typical behavioral issues monitors focus on when evaluating interviewers, we tabulated the specific codes used by the monitors as a whole, by the gold standard group, and by the active monitor group. We examined how each of the two monitor groups separately and both groups combined used the behavioral codes. By comparing the frequency distributions of each monitoring code, we were able to see if one group focused on a nonstandardized behavior more than the other group when evaluating the interviewers. To further examine consistency among monitors, we focused on whether or not monitors were consistent in documenting nonstandardized behaviors at the question level. We selected five short recordings and analyzed monitors' codes and feedback at the question level.

5. Results

The monitors' ratings of interviewers provided insight into our two research questions: (1) How accurate and consistent are monitors' overall ratings of interviewers? and (2) How accurate and consistent are monitors in identifying nonstandardized interviewer behaviors? In this section, we present our results, followed in the next session by a discussion of their implications.

5.1. How accurate and consistent are monitors' overall ratings of interviewers?

An examination of the overall ratings given by the three gold standard supervisors and 12 active monitors shows that the monitors are consistent in their use of the five-point ratings scale (Table 2). Consistent with our previous study's findings, the majority of ratings assigned were 2 (does not meet expectations) and 3 (meets expectations). Given the larger number of observations used for the current study, we were able to capture a small number of 1 (poor) and 4 (very good) ratings, but no ratings of 5 (excellent) were assigned.

Table 2: Overall Distribution of Ratings by Type of Monitor

Overall Rating	Gold Standard	Active Monitor	Overall
Poor (1)	0%	1%	1%
Does not meet expectations (2)	25%	29%	28%
Meets expectations (3)	73%	69%	70%
Very good (4)	2%	1%	1%
Excellent (5)	0%	0%	0%

To further explore monitors' consistency, we examined inter-rater reliability by tabulating monitors' ratings for each recorded interview (Table 3). We compared ratings among the gold standard monitors and the active monitors separately, as well as the overall agreement among all of the monitors. We found that the gold standard group was consistent, with 87 percent exact agreement. Both the active monitors' ratings and the overall ratings were 79 percent exact agreement—which leaves some room for improvement, considering that most of the assigned ratings were from only two out of five rating scale categories.

Table 3: Inter-Rater Agreement

Type of Monitor (n)	Observations	Percent Exact Agreement
Gold Standard (3)	60	87
Active Monitors (12)	240	79
All Monitors (15)	300	79

To determine the accuracy of the active monitors' overall ratings compared to those of the gold standard group, we compared the ratings of both groups and found a good degree of accuracy overall (Table 4). Nine of the 12 active monitors assigned the same rating as did the gold standard group for 70 percent or more of the interviews. The average level of accuracy between the active monitors and the gold standard monitors was 72 percent.

Table 4: Agreement Between Gold Standard and Active Monitors

Number of Active Monitors	Percent Agreement with Gold Standard Monitors
2	80
4	75
3	70
2	65
1	60

A new question we were able to explore with our current research was whether the monitors were consistent in their ratings over time. Nine of our monitors participated in the previous study, so we were able to see if there was any variation in their overall ratings by asking them to rate three interviews that they had rated the previous year. Twenty-five of the 27 ratings (93 percent) given in 2011 for these interviews were the same as those given in 2010. Only two ratings differed: one monitor assigned a 2 to an interview in 2010 but a 1 in 2011, another assigned a 2 to an interview in 2010 but a 3 in 2011 (Table 5).

Table 5: Variation Over Time

Monitors	2010 Rating	2011 Rating
Monitor #1	2	1
Monitor #2	2	3

5.2. How accurate and consistent are monitors in identifying nonstandarized interviewer behaviors?

To address the issue of consistency, we tabulated the amount of codes used for each of the key behavioral issues, both overall and for the two groups of monitors (Table 6). Across the 20 interviews evaluated by the 15 monitors, the most frequent type of comment made was general positive (49 percent), which is consistent with our findings from the previous study. In addition, 18 percent of the comments related to errors in asking questions (wording changes, skipping questions); 17 percent to probing issues (insufficient probing, leading, over-probing); 5 percent to feedback errors (inappropriate feedback, failure to provide feedback); 3 percent to general voice (volume, pace, clarity, tone) and rapport; and 6 percent to other nonstandard behaviors.

When comparing the frequency of codes assigned by active monitors to those of the gold standard group to determine their accuracy, there were few major differences. The probing and general nonstandard categories were the only codes in which we detected a large difference (differences of 11 and 16 percent, respectively). Upon investigating the difference in the general nonstandard coding, it was discovered that one gold standard monitor was using the code incorrectly, so this should be considered an anomaly.

Table 6: Behavioral Issues, Overall and by Study Group

Behavior Codes	All 15 Monitors (N=3023)	Three Gold Standard Monitors (N=546)	Twelve Active Monitors (N=2477)	Difference (GS-AM)
General Positive	49%	46%	50%	-4%
Question Asking	18%	20%	17%	3%
Probing	17%	8%	19%	-11%
Feedback	5%	3%	6%	-3%
Coding/Data Entry	3%	2%	3%	-1%
General Voice & Rapport	2%	2%	2%	0%
Other Nonstandard	6%	19%	3%	16%

We also wanted to know if the behavioral codes were consistent with the overall ratings. We looked at the behavioral codes by the overall rating to see if any interesting patterns emerged, but the results were predictable (Table 7). We focused on the interviews that were assigned an overall rating of 2 or 3, given that they comprised the majority of ratings. For example, there was a relationship between an interview's overall rating and the number of general positive comments given (the higher the rating, the more general positive codes were received). For the error codes, conversely, the lower the rating, the more error codes were assigned.

Table 7: Behavioral Codes by Overall Rating

Ratings	General Positive	Question Asking	Probing	General Voice & Rapport	Feed-back	Coding/Data Entry	Other Nonstandard
2 Rating	28	24%	25%	7%	8%	4%	4%
3 Rating	63%	14%	12%	5%	3%	3%	1%

Note: Percentages do not add to 100 percent due to rounding

Behavioral codes are assigned to particular questions; we wanted to know if monitors were assigning the same codes to the same questions. We examined four of the interviews at the question level (for a total of 337 questions asked) and found very little consistency in this area. Although there were 61 instances where two monitors used the same code for the same question, there was less consistency as more monitors came into play. There were 16 instances where three monitors coded a question identically, 10 instances where four monitors coded a question identically, and only 13 instances where five or more monitors coded a question identically.

6. Conclusion and Discussion

6.1 Conclusion

The goal of this research study was to explore monitors' behavior in order to improve the use of monitoring as a quality assurance tool. In particular, we focused on the consistency of monitors' ratings and the types of nonstandardized interviewing behavioral issues on which they focused. Based on an analysis of the data collected across 20 interviews evaluated by 15 monitors, we found that:

- Overall, the monitors were consistent in their use of the rating scale. The active monitors were consistent with one another, the gold standard monitors were consistent with one another, and the active monitors were consistent with the gold standard monitors. Furthermore, when the nine monitors who participated in the pilot study re-rated the same interviews, their ratings over time were consistent.

- While the overall consistency among all monitors was good (79 percent overall), the level of exact agreement could be improved, especially considering that monitors assigned ratings across only two rating scale levels. While four of the five levels on the rating scale were used by the monitors, the majority of their ratings were at levels 2 (does not meet expectations) and 3(meets expectations).

- The monitors' use of nonstandardized behavioral codes was consistent with the overall ratings they assigned to interviews. While the active and gold standards monitors tended to use the behavior codes in the same proportion, active monitors flagged "probing" issues more frequently than did the gold standards group. Although monitors did not always comment on the same question, when they did, they tended to assign the same behavioral codes.

6.2 Discussion

In exploring the consistency of monitors' use of the rating scale, we learned that the monitors were very consistent but did not use the full rating scale. This raises the question of whether our five-level scale should be revised or replaced. If the only way to achieve a rating of 4 (very good) is when the interview is challenging, and a 1 (unacceptable) or a 5 (excellent) is rarely assigned, is the 1–5 scale really useful?
If we replaced the numbered scale with feedback statements (that is, "Needs immediate attention," "Needs extensive retraining," "Needs retraining in one or two areas," "No issues, excellent job"), would monitors be more willing to use the full range?

We found it interesting that, although monitors often differed in which parts of the interview they flagged with behavioral codes as problematic or noteworthy, they also often arrived at the same overall rating. This leads us to question the relationship between the overall ratings and the behavioral codes, and the underlying purpose of monitoring. If the main purpose of interviewer monitoring is as a quality control mechanism—to identify and correct errors in data collection and ensure consistency across interviewers—then monitoring interviews in real time and assigning consistent behavior codes is important.

On the other hand, if another purpose of monitoring is to improve the skills of the interviewers, then the assignment of behavior codes is guided by professional development needs rather than data quality needs. Interviewers and monitors can

listen to taped interviews, and behavior codes can be used to identify areas where the interviewer needs further training. To improve interviewers' overall skills, monitors might focus on a few types of errors at a time to reinforce the need for improvement in these areas.

In examining monitors' consistency across time, we only had access to nine monitors' ratings of three interviewers from prior monitoring cycles. Because consistency across time is an important indicator of the quality of monitoring, further study of monitors' consistency across time, with larger groups of monitors and a larger sample of interviews, would provide valuable information about variations in monitors' performance from project to project and within projects across time.

While the current study included a wide variety of projects and lengths of interviews, we did not systematically select interviews from key points in the projects' life cycles. For projects with long data collection periods, interviews conducted at the beginning include learning time for the interviewers and monitors. In the middle of a project, interviewers and monitors are usually performing at their best, while toward the end, fatigue may affect their skills and performance. To provide a more robust assessment of monitors' consistency, future studies could sample interviews from the beginning, middle, and end of projects' life cycles.

To further explore monitor quality, it would be interesting to compare monitoring accuracy and consistency by experience level (seasoned versus novice monitors), and also by varying the experience level of the interviewers monitored (seasoned versus novice interviewers).

Acknowledgements

We wish to give special thanks to the Building Strong Families (BSF), the Evaluation of Individual Training Account Demonstration (ITA), the Early Head Start Family and Child Experiences Survey (BabyFACES), the Community Tracking Study (CTS), the National Beneficiary Study (NBS), and the National Mental Health Services Survey (N-MHSS) projects for use of their recorded interviews. We would also like to thank Ron Palanca for providing SAS programming, and the Survey Operations Center Monitoring staff for their participation and cooperation.

References

Baker, J., C. Gentile, J. Markesich, and S. Marsh. 2010. Who's Monitoring the Monitors? Examining Monitors' Accuracy and Consistency to Improve the Quality of Interviews. In *JSM Proceedings.* Alexandria, VA: American Statistical Association.

Burks, A. T., P. J. Lavrakas, K. Steve, K. Brown, B. Hoover, J. Sherman, and R. Wang. 2006. How organizations monitor the quality of work performed by their telephone interviewers. In *Proceedings of the Survey Research Methods Section*, American Statistical Association, 4047-4054.

Cannell, C., and L. Oksenberg. 1988. Observation of behavior in telephone interviews. In, *Telephone Survey Methodology*, R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L Nicholls, and J. Waksberg (eds.). New York: John Wiley and Sons Inc. 475-495.

Couper, M. P., L. Holland, and R. M. Groves. 1992. Developing systematic procedures for monitoring in a centralized telephone facility. In *Journal of Official Statistics, 8*(1), 63-76.

Currivan, D., E. Dean, and L. Thalji. 2006. Using standardized interviewing principles to improve a telephone interviewer monitoring protocol. Presented at the 2nd International Conference on Telephone Survey Methodology, Miami, FL.

Durand, C. 2005. Measuring interviewer performance in telephone surveys. In *Quality and Quantity, 39*(6), 763-778.

Fowler, F.J., and T. J. Mangione. 1990. *Standardized survey interviewing: Minimizing interviewer-related error*. Newbury Park, CA: Sage Publications.

Lavrakas, P. J. 2010. Telephone surveys. In *Handbook of survey research*, P. V. Marsden and J. D. Wright (eds.). London: Emerald Group Publishing, Limited. 471-498.

Mudryk, W., M. J. Burgess, and P. Xiao. 1996. Quality control of CATI operations in Statistics Canada. In *Proceedings of the Survey Research Methods Section*, American Statistical Association. 150-159.

Steve, K. W., A. T. Burks, P. J. Lavrakas, K. D. Brown, and J. B. Hoover. 2008. Monitoring telephone interviewer performance. In *Advances in telephone survey methodology*, J. M. Lepkowski, C. Tucker, J. M. Brick, E. D. de Leeuw, L. Japec, P. J. Lavrakas, M. W. Link, and R. L Sangster (eds.). New York: John Wiley and Sons Inc. 401-422,

Sudman, S. 1967. Quantifying interviewer quality. In *Public Opinion Quarterly*. *30*(4), 664-667.

Tarnai, J. 2007. Monitoring CATI interviewers. Presented at the 62nd Annual Conference, American Association of Public Opinion Research, Anaheim, CA.