Who's Monitoring the Monitors? Examining Monitors' Accuracy and Consistency to Improve the Quality of Interviews

Joseph Baker, Claudia Gentile, Jason Markesich
Shawn Marsh
Mathematica Policy Research, Inc., 600 Alexander Park, Princeton, NJ 08543-2393

Abstract

Most survey organizations use monitoring to evaluate the performance of telephone interviewers, identify problems with survey questions, and provide recommendations for interviewing techniques or methodological investigations of questionnaire designs. Though monitoring is viewed as a key quality assurance tool, little research has been devoted to understanding the behavior of monitors—specifically, their ability to provide effective and consistent feedback on interviewer performance. To explore monitors' behavior, Mathematica Policy Research designed a monitoring consistency exercise that addressed the following questions: (1) What typical behavioral issues do monitors focus on when evaluating interviewers? (2) What criteria do monitors use to rate interviewers? (3) Are monitors coding non-standardized interviewer behavior accurately and consistently? and (4) What is the extent of monitor variation within each monitor (drift across monitoring sessions) and between monitors (differences in leniency and severity between different monitors)?

For this exercise, we recruited two groups of monitors: three monitoring supervisors, who served as the gold standard, and eight active monitors, with a range of monitoring experience. The monitors evaluated eight digitally recorded interviews using a five-point Interviewer Rating Scale and a behavioral coding system that captures both positive and non-standardized interviewer behavior. To explore the ways monitors use criteria to assign ratings, we conducted focus group discussions in which monitors discussed their processes and decision making, revealing how they apply the criteria when assigning ratings and providing feedback to interviewers.

Analyses of monitors' overall ratings of interviewers, the specific behavior issues they include in feedback to interviewers, and their explanations of their ratings revealed that both the gold standard and active monitors were very consistent in their overall ratings. However, there was little variation in their ratings; monitors assigned only ratings of "2" (does not meet expectations) and "3" (meets expectations). Almost half of monitors' feedback comments were positive; among the non-standardized behavior issues noted by monitors, about 40 percent related to probing errors. When rating interviewers, monitors used criteria that related to interviewers' performance (as articulated in the Interview Rating Scale), but also used criteria that were not part of the Interview Rating Scale, such as the interviewer's experience level, past performance, and familiarity with the project.

Key Words: monitoring, inter-rater reliability, assessing interviewer performance, consistency training, quality control, data quality

1. Introduction

Like most survey research organizations, Mathematica Policy Research monitors telephone interviewers to ensure that they follow standardized interviewing procedures so that the data they collect is of high quality. In addition, we use monitoring 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 non-standardized 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 monitors use different criteria, or criteria that are not part of the rating scale, when scoring interviewing sessions, interviewer evaluation scores might be inaccurate.
- 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 focused on describing monitoring processes or methods, such as the key elements of an effective monitoring system (Cannell & Oksenberg, 1988; Fowler & 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).

Thus, little is known from research about the factors that affect 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:

- What typical behavioral issues do monitors focus on when evaluating interviewers?
- What criteria do monitors use to rate interviewers?
- Are monitors coding non-standardized interviewing behavior accurately and consistently?
- What is the extent of monitor variation, within each monitor and between monitors?

As part of this exercise, we asked two groups of monitors (gold standard and active monitors) to evaluate eight digitally recorded interviews from three telephone surveys. We also conducted focus group discussions with each group of monitors, to explore the ways monitors use criteria to assign ratings. This paper presents the key findings of this monitoring consistency exercise. 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 only an introduction, a complete interview, or a partial interview.
- **A 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, re-reading 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, re-reading 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 re-reading 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 February and March 2010. In this section, we describe the subjects and materials used to carry out the study, the data collection procedures, and the methods of analysis.

3.1 Subjects

Eight active monitors and three supervisors were recruited for this study to evaluate eight digitally recorded telephone interviews conducted by a cross-section of interviewing staff. The eight active monitors recruited for this study represent a range of experience, with 2 to 16 years of experience interviewing and monitoring. The three supervisors had

10 to 22 years of experience interviewing and monitoring, and 5 to 20 years of experience as supervisors. They served as the gold standard and 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); and (3) The Early Head Start Family and Child Experiences Survey (EHS) (parents interviewed about their children's experiences with the EHS program).

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 classified 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). In addition, to ensure a full range of skill levels, we included novice interviewers. We then randomly selected two above-average interviewers, three average and three below-average interviewers from the pool of interviewers engaged in the three projects mentioned above.

Because monitors evaluate both complete and partial interviews, we included three complete interviews and five partial interviews, each of which contained an introduction. Lastly, we selected one digital recording from each of the eight interviewers, taking into consideration the need to select a mix of complete and partial sessions. None of the digital recordings selected for the study had been previously evaluated by a monitor or supervisor. Table 1 provides a summary of the selected recordings.

Table 1: Summary of Recorded Interviews

Project	Interviewer Skill Level	Session Type	Length of Time (Minutes: Seconds)
BSF	Above Average	Complete	22:51
ITA	Above Average	Complete	17:21
ITA	Average	Complete	21:29
BSF	Average	Partial	15:00
BSF	Average	Partial	12:00
EHS	Below Average	Partial	15:00
ITA	Below Average	Partial	15:00
EHS	Below Average	Partial	10:44

3.3 Data Collection

To carry out the monitoring consistency exercise, during a two-week period we scheduled individual monitoring sessions with each study group: the eight 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 process and form. We also informed the participants that they would monitor and evaluate the digitally recorded interviews independently of each other, 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 non-standardized interviewing behaviors and positive aspects of the interviews in a monitoring database. They each evaluated the eight digital recordings, yielding a total of 88 observations for the study.

To better explore how monitors use the evaluation criteria to assign ratings, we conducted focus group discussions with members of the gold standard group after they evaluated each recording. During the focus group discussions, we asked each of the monitors to discuss their overall rating, what they considered to be key issues that surfaced during the interview, and the single most important non-standardized or positive behavior that they thought dictated the overall score. We also asked the group to assign alternate ratings that are not part of the current evaluation form, such as letter grades (A–F) or a adding a plus (+) or minus (-) sign to the numerical rating. In addition, we conducted one focus group with a subset of the active monitors, to gain additional perspective on their decision-making processes, including how they apply the criteria when assigning ratings and providing feedback.

3.4 Data Analysis

To address the question of what typical behavioral issues monitors focus on when evaluating interviewers, we tabulated the specific codes used by the all the monitors, by the gold standard group, and by the active monitor group. We examined how each of the two monitor groups and both groups used the behavioral codes. By comparing the frequency distributions of each monitoring code, we were able to see if one group focused on a non-standardized behavior more than the other group when evaluating the interviewers.

To address the question of what criteria monitors use when rating interviews, we analyzed the focus group discussions (field notes and audio tapes of these discussions) to identify key issues. We shared these issues with the gold standard team to gain confirmation. To address the question of how accurate and consistent monitors were in

their ratings of interviews, we analyzed the inter-rater agreement among all monitors and within each group (gold standard and active). To address the question of the extent of monitor variation within and between monitors, we examined the inter-rater agreement data and patterns in monitors' coding of non-standardized behavior issues.

4. Results

The rating of interviewers by gold standard and active monitors and the focus group discussions with these two groups yielded useful information related to the four research questions: (1) What typical behavioral issues do monitors focus on when evaluating interviewers? (2) What criteria do monitors use to rate interviewers? (3) Are monitors coding non-standardized interviewing behavior accurately and consistently? and (4) What is the extent of monitor variation, within and between monitors? In this section, we present the results, followed by a discussion of their implications in section 5.

4.1 What Typical Behavioral Issues Do Monitors Focus on When Evaluating Interviews?

To address this question, we tabulated the monitors' ratings across the eight interviewers rated by each monitor. Table 2 presents the percentage of comments made about each of the key behavior issues, both overall and for the two study groups: gold standard and active monitor. Across the eight interview sessions evaluated by the 11 monitors, 22 percent of the comments related to probing issues (i.e., insufficient probing, leading, over-probing); 15 percent to errors in asking questions (i.e., wording changes, skipping questions); 9 percent to feedback issues (i.e., inappropriate feedback, failure to provide feedback); 5 percent to coding/data entry errors (i.e., incorrect entry or coding); and only 3 percent to general voice (i.e., volume, pace, clarity, tone) and rapport. The most frequent type of comment made was General Positive. (43 percent); only 3 percent were other nonstandard behaviors.

In comparing the gold standard monitors with the active monitors, the identification of behavior issues is very similar. However, although a high percentage of the comments made by both groups were General Positive, almost half of the active monitors' comments (47 percent) were General Positive, with only one-third of the gold standard monitors' comments (35 percent) General Positive. Also, the gold standard monitors commented on probing and question-asking issues slightly more often than did the active monitors (6 and 8 percent more, respectively).

Table 2: Behavioral Issues, Overall and by Study Group

Behavior Issues	All Eleven Monitors (N = 790)	Three Gold Standard Monitors (N = 247)	Eight Active Monitors (N = 543)	Difference (GS-AM)
Probing	22%	26%	20%	6%
Question Asking	15%	20%	12%	8%
Feedback	9%	6%	10%	-4%
Coding/Data Entry Error	5%	7%	5%	2%
General Voice & Rapport	3%	2%	3%	-1%
General Positive	43%	35%	47%	-12%
Other Nonstandard	3%	4%	3%	1%

When we examined the number of times monitors noted specific behavior issues (see Table 3), clear patterns of emphasis emerged. Although respondents used almost all of the nonstandard behavior categories, they used certain categories more frequently. Across all the monitors, almost one-fifth of the 421 comments about non-standardized behavior issues related to "insufficient probe or failure to probe," one-fifth to "leading or evaluative probe," and one-fifth to "major wording change during question asking." The rest of the comments were distributed across the categories.

For the gold standard and active monitors, these three categories also received a high percentage of the comments. In comparing the gold standard monitors with the active monitors, the general pattern of issues is very similar. However, the active monitors commented more often on "leading or evaluative probes" and "inappropriate feedback" than did the gold standard group.

Table 3: Non-standardized Behavioral Categories, Overall and by Study Group

Non-standardized Behavior Categories	**All Eleven Monitors** *% of Comments* **(N = 421)**	**Three Gold Standard Monitors** *% of Comments* **(N = 151)**	**Eight Active Monitors** *% of Comments* **(N = 270)**	**Difference (GS-AM)**
SPECIFIC BEHAVIOR CATEGORY: PROBING				
Major wording change	3	5	2	3
Insufficient/failure to probe	18	20	16	4
Leading or evaluative probe	18	14	21	-7
Inappropriate definition	0*	0	0*	0
Over-probing	1	3	0*	3
Other error	0*	0	0*	0
SPECIFIC BEHAVIOR CATEGORY: QUESTION ASKING				
Major wording change	19	22	18	4
Skipped question	2	3	1	2
Other error	6	8	5	3
SPECIFIC BEHAVIOR CATEGORY: FEEDBACK				
Inappropriate feedback	12	8	15	-7
Failure to give feedback	2	1	2	-1
Other error	3	2	3	-1
SPECIFIC BEHAVIOR CATEGORY: CODING DATA/ENTRY ERROR				
Incorrect entry	7	9	6	3
Misuse of CATI conventions	0*	1	0	1
Other error	2	1	3	-2
SPECIFIC BEHAVIOR CATEGORY: GENERAL VOICE AND RAPPORT				
Volume	0	0	0	0
Pace	3	3	3	0
Clarity	1	0	1	-1
Tone	0*	0	1	-1
Rapport	1	1	2	-1

[1] Percentages do not add to 100 percent due to rounding.
*Fewer than 1 percent of monitors used this category.

In addition to these differences in types of behavior issues noted by the two monitoring groups, there were slight differences in the types of behavior issues to which monitors called attention, relative to the overall ratings of the interview session. Table 4 presents the percentage of behavior issues noted by all of the monitors for interviewers rated 2 (does not meet expectations) and interviewers rated 3 (meets expectations). As expected, interviewers rated 3 received a higher proportion of positive comments and fewer comments about behavior issues, with one exception (coding/data entry error).

Table 4: Differences in Ratings by Key Issues for Interviews Rated 2 (Does Not Meet Expectations) and 3 (Meets Expectations)

Behavior Issues	**Interviews Rated 2** (N = 346)	**Interviews Rated 3** (N = 444)	**Difference** (2–3)
Probing	27%	17%	10%
Question Asking	22%	9%	13%
Feedback	15%	4%	11%
Coding/Data Entry Error	4%	7%	-3%
General Voice and Rapport	5%	1%	4%
General Positive	22%	60%	38%
Other Nonstandard	5%	2%	3%

4.2 What Criteria Do Monitors Use to Rate Interviewers? What Factors Influence Monitors' Ratings?

From the focus group discussions with the gold standard and active monitors, five key factors emerged that influenced their ratings and the type of feedback they provided to interviewers (see Figure 2 below).

Monitors' familiarity with project. Both the gold standard and active monitors often began the explanation of their ratings by stating how familiar they were with the project for which the interview was conducted. Familiarity with the project gave them confidence in their evaluation and a context for understanding the conventions and project-specific expectations of interviews. When unfamiliar with the project, monitors reported relying on their knowledge of basic interviewing skills, which they claimed were the core of interviewer training across all projects.

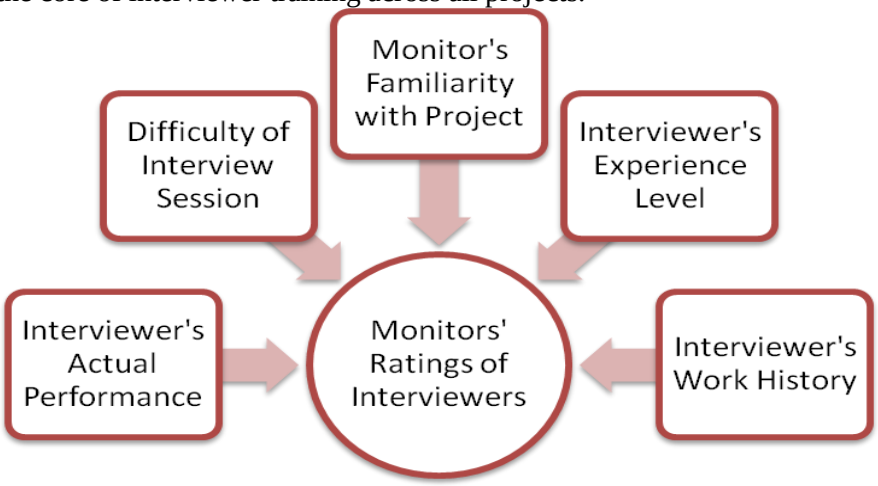

Figure 2: Factors Influencing Monitors' Ratings of the Quality of Interviewers

Interviewer's work history. The gold standard team and active monitors frequently discussed the interviewer's work history. Before conducting a new monitoring session, monitors said that they typically review the monitoring reports to learn about the interviewer's performance. If an interviewer makes a few errors, but the types of errors are the same as those noted in previous monitoring sessions, the monitor considers this continual pattern of errors as more serious than simply the number of errors made in the session being monitored. Some monitors reported assigning a rating of 2 (does not meet expectations) if the interviewer continues to make the same mistakes.

Interviewer's experience level. Likewise, both groups of monitors reported treating errors made by experienced interviewers as more serious than those made by novice interviewers (with the exception of two active monitors). They defined novice interviewers as those who had fewer than two weeks of experience conducting interviews for a particular project. In particular, the gold standard team agreed that they tended to assign a rating of 3 (meets expectation) rather than a 2 (does not meet expectation) to novice interviewers because they did not want to discourage the novice. They would record all the errors made and discuss these with the novice, but they felt that to give a novice a rating of a 2 would be too severe. If experienced interviewers, who "should have known better," made the same serious errors, the gold standard team would give them a 2 to signal the need for retraining. Thus, most of the monitors use the rating system as a way to coach and encourage interviewers rather than as a purely evaluative tool.

Difficulty of the interview session. All of the study participants reported that they consider the difficulty of the interview session when assigning ratings. The only way an interviewer was assigned a rating of 4 (very good) was if the interview session proved especially challenging. None of the study participants reported ever assigning a rating of 5 (excellent). The types of challenges that might lead to a rating of 4 or 5 are:

- Completing an especially long and/or complicated interview,
- Converting a reluctant respondent,
- Remaining professional with a belligerent or hostile respondent,
- Persevering with a respondent with a physical or mental disability (such as someone who is hard of hearing, has an apparent cognitive or developmental disability, or a speech impediment).

Interviewer's actual performance. Both groups of monitors reported basing their ratings on the interviewers' actual performance. They said they based their ratings on the number of errors, types of errors, and an overall sense of whether the interviewer had obtained accurate data for the project. Both groups almost always considered recording inaccurate information a serious error. Two of the most common serious errors were (1) poor listening, leading to miscoding, which sometimes also led to the wrong series of follow-up questions; and (2) lack of probing, resulting in incomplete responses and/or missed opportunities to collect in-depth information.

Both groups of monitors reported rarely assigning a rating of 1 (Unacceptable), and only when experienced interviewers made "unforgivable" errors, such as falsifying data. When they hear interviewers making "unforgivable" errors during live monitoring sessions, monitors interrupt the interview and provide guidance so that the interviewers can remedy their mistakes, thus avoiding a rating of 1.

Additional lessons learned from the focus group discussions. In addition to the information gained about the types of criteria monitors use to evaluate and provide feedback to interviewers, the focus group discussions yielded several insights about how monitors view their role and what they value:

- Monitors highly value obtaining good quality, accurate data. Poor data is often a reason for giving a low rating.
- Monitors provide support and corrective guidance to interviewers even when they are not actively monitoring.
- Monitors highly value interviewers who can convert refusals, thus improving response rates.
- Monitors working the same shift often consult each other about the monitoring process to ensure fairness and consistency in their feedback to interviewers.
- Monitors are concerned about the impact of their feedback on staff retention.
- When monitors were asked to use a grading scale (A to F) and a scale that includes pluses and minuses, their ratings were equally consistent as when they used the original 1 to 5 scale.

4.3. Are Monitors Rating Non-standardized Interviewing Behavior Accurately and Consistently?

To examine inter-rater reliability, we tabulated monitors' ratings for each recorded interview. We compared ratings among the gold standard monitors, among the active monitors, and the overall agreement among all of the monitors (Table 5). In general, monitors achieved a high level of agreement when assigning the overall ratings and most of the disagreement could be traced to two active monitors. This level of agreement is not surprising, given that monitors in both groups did not use the full rating scale and only assigned ratings of 2 (Does Not Meet Expectations) or 3 (Meets Expectations).

An examination of the percentage of exact agreement among monitors by type of interview shows that the gold standard group achieved a very high level of agreement across all the interviews, with only one gold standard monitor disagreeing about one interview rating (one complete interview). The active monitors also achieved a good level of agreement: all eight agreed on one complete and two partial interviews; seven of eight agreed on one complete and one partial interview; and six of eight agreed on one complete and two partial interviews.

It is important to note that our analysis of rater consistency is limited by the small number of interviews and the lack of balance among the type of interviews and interviewer skill level (i.e., the complete interviews had only above average and average interviewers, and the partial interviews had only average and below average interviewers). Likewise, Table 5 reports only the percentage of exact agreement. We did not conduct further analyses due to the limited range of ratings assigned by monitors. Although the Interviewer Rating Scale (see Figure 1) consists of five levels of performance, the monitors in this study assigned only ratings of 2 or 3, effectively reducing the scale to a binary rating system.

Table 5: Percentage of Exact Agreement, by Interview and Overall

Type of Interview	Interviewer Skill Level	Percentage Agreement: 3 Gold Standard Monitors	Percentage Agreement: 8 Active Monitors	Percentage Agreement: All 11 Monitors
Complete	Above Average	100	100	100
Complete	Above Average	67	75	73
Complete	Average	100	88	91
Partial	Average	100	100	100
Partial	Average	100	88	91
Partial	Below Average	100	100	100
Partial	Below Average	100	75	82
Partial	Below Average	100	75	82
TOTAL		96	88	90

4.4 What Is the Extent of Variation Within and Between Monitors?

In examining further the degree of accuracy and consistency among monitors, we sought to explore (1) the within-monitor variation, to assess whether individual monitors drift in their assignment of ratings, becoming more lenient or severe across monitoring sessions; and (2) patterns in between-monitor variation, to assess whether individual monitors are more lenient or more severe compared with other monitors.

However, our examination of these issues was limited by the number of interview sessions in this study: eight. In order to evaluate drift and degrees of leniency or severity, monitor's ratings across a stretch of time (at least one week) are necessary. The eight interview sessions evaluated for this study represent less than one day's monitoring activities. Thus, the time it took to monitor eight interview sessions did not provide us with a long enough period over which to detect patterns of leniency and/or severity, both within and among monitors. In addition, the monitors' use of a limited range on the rating scale also limited our ability to detect patterns of leniency and/or severity.

5. Conclusion and Discussion

5.1 Conclusion

Although monitoring is a critical quality assurance tool, little research has been conducted about the behaviors of monitors. The goal of this research study was to shed some light on monitors' behavior, in particular, the types of non-standardized interviewing behavioral issues on which monitors focus, the criteria they use to evaluate and rate interviewers, and the consistency of their ratings. Based on an analysis of the data collected across eight interviews evaluated by 11 monitors as well as focus group discussions with monitors, we found that:

- Almost half of the monitoring codes and comments across all of the observations were positive, indicating that monitors value positive feedback when evaluating interviewers.

- Approximately 40 percent of the behavior issues that monitors commented on were related to probing errors, more than any other type of non-standardized behavior. This was true for interviewers rated 2 (does not meet expectations) and 3 (meets expectations).
- The length of the interview was not related to the degree of agreement among monitors; shorter and longer interviews were rated with the same high degree of consistency.
- Although the monitors achieved a high level of agreement when scoring the interviewers, there was little variation in terms of the ratings assigned to the interviewers (only ratings of 2 or 3 were assigned).
- Monitors used criteria about interviewer's performance from that the Interview Rating Scale when assigning ratings. However, they also used criteria not found in the scale, such as the interviewer's familiarity with project, past performance on the project, and experience level.
- Monitors defined their role broadly, to encompass achieving high data quality across all projects by developing and maintaining stable, experienced interview staff.

5.2 Limitations and Discussion

This study was intended to be an exploration of monitors' accuracy and consistency. As such, the scope of this study was limited to 11 monitor's ratings of eight interviews and focus group discussions with the gold standard and active monitors. The analysis yielded useful information about how monitors use the rating scale and the behavioral codes, their overall consistency, and the criteria they use to assign ratings. However, we did not have sufficient evidence to explore issues of individual monitor drift and patterns of leniency and severity. Further studies that explore monitoring across several weeks, instead of several days, would provide more in-depth insight into long-term monitor behavior.

In addition, we did not collect information from the interviewers about their views of and experiences with the monitoring system. This unexplored area could be the focus of future studies to provide insight about questions such as the following:

- When monitored, do interviewers focus on their overall rating or the specific feedback on behavioral issues, or on both?
- Does monitoring help interviewers improve?
- What is the impact of monitoring on staff development and retention?

The findings from this study also raise several questions, the answers to which will help us improve our monitoring system and the training and supervision of monitors.

- 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 (i.e., "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?
- If monitors use ratings of 2 (does not meet expectations) and 3 (meets expectations) differently for experienced and novice interviewers, are these ratings more a communication tool than an evaluation tool? Should we adjust our training

procedures to ensure the consistent application of ratings, independent of the degree of interviewers' experience?

- Can the monitoring system be adjusted to provide for a way to examine the consistency in behavioral coding at the question level, to see whether monitors agree not only about the overall rating, but also about the interview behaviors they identify as needing improvement?

Acknowledgements

We wish to give special thanks to the Building Strong Families (BSF), the Evaluation of Individual Training Account Demonstration (ITA), and the Early Head Start (EHS) Family and Child Experiences Survey projects for use of their recorded interviews and to those individuals whose help is greatly appreciated, including Jackie Donath, Hugo Andrade, Beverly Kelly, Pat Ubriaco, Karen Groesbeck, Marianne Stevenson and the Survey Operations Center Monitoring staff.

References

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

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

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

Currivan, D., Dean, E., & Thalji, L. 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. *Quality and Quantity*, 39(6), 763-778.

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

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

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

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

Sudman, S. 1967. Quantifying interviewer quality. *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.