MATHEMATICA Policy Research

ISSUE BRIEF

MAY 2013

Who's Monitoring the Monitors? Ensuring Data Quality in Telephone Interviewing

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

Survey organizations monitor interviewers to ensure data quality, yet little is known about the factors that affect monitors' judgments and the feedback they provide to interviewers. Because monitoring is a critical component of quality assurance, it is important to understand monitors' behaviors—specifically, their ability to provide effective and consistent feedback on interviewers' performance. In this issue brief, we summarize the findings of a study of monitors' behaviors conducted by Mathematica Policy Research, as well as a summary of the literature on monitoring.

What Does the Literature Show?

Research on understanding monitors' behavior is not extensive. Most studies have described monitoring processes or methods, such as 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). In 2010, Mathematica conducted an exploratory study of monitors' consistency and accuracy, indicating a need for a more in-depth examination of monitors' behaviors and the factors that influence their judgments (Baker et al. 2010).

Gauging Monitors' Consistency and Accuracy

In a 2011 study (Baker et al. 2011), researchers at Mathematica explored the consistency and accuracy of two groups of monitors: 3 monitor supervisors who had from 5 to 15 years of experience monitoring and supervising staff and 12 active monitors who had from one to 17 years of experience interviewing and monitoring. Both groups evaluated the same 20 digitally recorded interviews from six projects (15 complete interviews, lasting from 10 to 50 minutes each, and 5 partial interviews, lasting from 10 to 40 minutes each). The 20 interviewers included those whose past ratings were below average, average, or above average.

In addition, to gauge within-monitor consistency, nine monitors who had rated interviews in the 2010 study were assigned the same three interviews to rate in 2011. During the monitoring session, monitors first evaluated interviews using the behavioral codes listed in Figure 1 and then assigned an overall rating for each session using a five-point evaluation scale (Figure 1).

What We Found: Consistency and Accuracy of Overall Ratings

The overall consistency of ratings was close to the target level of 80 percent agreement. The 3 monitor supervisors assigned the same overall ratings for 87 percent of the interviews and the 12 active monitors for 79 percent of the interviews. Combined, the overall agreement rate was 79 percent.

Figure 1. Mathematica Interview Monitoring Codes and Rating Scale

Behavioral Code	Overall Evaluation Scale
1. Question-asking errors	1. Unacceptable
2. Probing errors	2. Does not meet expectations
3. Feedback errors	3. Meets expectations
4. Coding or data entry errors	4. Very good
5. Voice and rapport errors	5. Excellent
6. Positive comments	

Monitors were consistent not only with one another, but with themselves. Of the nine monitors who rerated three interviews, seven gave the same rating both years and two gave ratings within one level of their original ratings.

To assess the accuracy of the ratings, we compared the 12 active monitors' ratings with those of the 3 monitor supervisors; we found that the active monitors assigned the same rating as the monitor supervisors for 72 percent of the interviews monitored.

What We Found: Range and Consistency of Behavioral Codes

Monitors commented on a range of interviewer behaviors; almost half of the comments were generally positive (Figure 2). Overall, the 3 monitor supervisors and the 12 active monitors commented the same number of times on the different behavioral issues, with two exceptions: active monitors commented more on probing issues than did the monitor supervisors, and the monitor supervisors commented more on other nonstandard behaviors than did the active monitors. However, when we compared monitors' behavioral codes by interview, we found very little consistency. When we looked across the 20 interviews, monitors seemed to comment on the same behaviors, but they rarely commented on the same behaviors on the same interview.

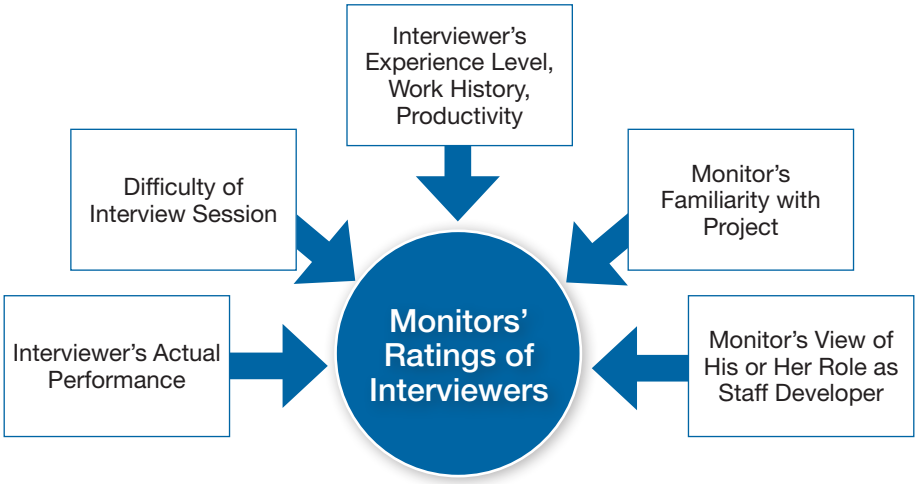
What We Found: Factors Affecting Monitors' Judgments

To better understand the factors that affect monitors' judgments, we conducted two focus groups: one with the monitor supervisors and the other with the active monitors. We asked them to discuss how they arrived at their overall ratings and what key issues surfaced during the interviews. These discussions revealed that, in addition to interviewers' actual performance, monitors took into account four other factors not included in the evaluation scale (Figure 3).

Figure 2. Mostly Frequently Used Behavior Codes

Behavior Code	Percentage of Time Used		
	All Monitors	Active Monitors	Monitor Supervisors
Positive Comments	49	50	46
Question-Asking Errors	18	17	20
Probing Errors	17	19	8
Feedback Errors	5	6	3
Coding or Data Entry Errors	3	3	2
Voice and Rapport Errors	2	2	2
Other Errors	6	3	19

Figure 3. Factors Affecting Monitor's Judgments



What We Found: Interviewers' Perspectives on Monitoring

To address the issue of whether monitors provided consistent feedback to interviewers and to obtain insight into interviewers' experiences with the monitoring system, we conducted focus group discussions with new and experienced interviewers. These discussions indicated the following about the interviewers:

- They found the monitoring sessions helpful.
- Interviewers felt that most monitors tried to focus on the positive aspects of the interview.
- They found variation in monitors' feedback, in that some monitors were stricter whereas others were more lenient.
- They felt that differences in monitors' communication styles and skills affected the usefulness of their feedback.
- Interviewers noted that feedback sessions were most useful when they received feedback immediately after the interview.
- They paid more attention to feedback on the behavioral codes (positive and negative comments) than on their overall ratings.

Main Lessons Learned

Consistency and accuracy. Monitors are basically consistent and accurate in their use of the rating scale. However, interviewers note that some monitors were stricter than others, indicating a need to retrain monitors to enhance their consistency and accuracy.

Range and consistency of behavioral codes. Although monitors employ a range of behavioral codes, they are not consistent in assigning these codes, indicating a need to revise the behavioral codes and/or the process monitors follow to improve their consistency.

Factors affecting monitors' judgment. When assigning ratings, monitors considered factors beyond the interviewers' actual performance, indicating a need to revise the monitoring system to clarify the key evaluation dimensions and allow for the addition of new dimensions or factors.

Interviewers' perspectives on monitoring. Interviewers found the overall system helpful but noted that variations in communication styles and degree of severity among monitors affected the usefulness of the monitoring session, indicating a need for further monitor training to enhance communication skills and consistency.

Looking Ahead

Based on our findings, we have several recommendations:

1. When examining the issue of monitor consistency, it is helpful to look beneath the surface. Using exercises—such as having monitors evaluate and discuss the same interviews—is an effective way to explore monitors' decision making and the criteria they use. These criteria can then be compared with any rating scales that the monitors are expected to use, to see whether they are focusing on elements of the interview consistent with the criteria in the rating scales.

2. If necessary, alter rating scales to ensure consistency across monitors. If the monitors note the criteria they use to rate interviewers that are not already included in the rating scale, consider revising the scale to include these criteria or find another way to standardize the rating process so that all monitors consider the same criteria when rating interviewers.

3. Provide training to monitors on how to provide feedback. Some monitors will naturally be better than others at delivering feedback to interviewers. However, all monitors can be trained to provide clear and constructive feedback. Providing monitors with a framework on how to deliver feedback and with training opportunities to learn and practice these skills will result in more consistent and useful feedback to the interviewers.

References

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

Baker, J. C. Gentile, J. Markesich, S. Marsh, and R. Weiner. "Ensuring Data Quality: Monitoring Accuracy and Consistency Among Telephone Interview Monitors." *JSM Proceedings*. Alexandria, VA: American Statistical Association, 2011.

Burks, A. T., P. J. Lavrakas, K. Steve, K. Brown, B. Hoover, J. Sherman, and R. Wang. "How Organizations Monitor the Quality of Work Performed by Their Telephone Interviewers." *Proceedings of the Survey Research Methods Section*, American Statistical Association, 2006. 4047-4054.

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

Couper, M. P., L. Holland, and R. M. Groves. "Developing Systematic Procedures for Monitoring in a Centralized Telephone Facility." *Journal of Official Statistics, 8*(1), 1992. 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, 2006.

Durand, C. 2005. "Measuring Interviewer Performance in Telephone Surveys. *Quality and Quantity, 39*(6), 2005. 763-778.

Fowler, F. J., and T. J. Mangione. *Standardized Survey Interviewing: Minimizing Interviewer-Related Error.* Newbury Park, CA: Sage Publications, 1990.

Lavrakas, P. J. "Telephone Surveys." In *Handbook of Survey Research.* Eds. P. V. Marsden and J. D. Wright. London: Emerald Group Publishing, Limited, 2010. 471-498.

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

Steve, K. W., A. T. Burks, P. J. Lavrakas, K. D. Brown, and J. B. Hoover. "Monitoring Telephone Interviewer Performance." In *Advances in Telephone Survey Methodology.*" Eds. J. M. Lepkowski, C. Tucker, J. M. Brick, E. D. de Leeuw, L. Japec, P. J. Lavrakas, M. W. Link, and R. L Sangster. New York: John Wiley and Sons Inc., 2008. 401-422.

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

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