Who's Monitoring the Monitors? Examining Monitors' Accuracy and Consistency to Improve the Quality Assurance Process

May 16th, 2010
Presentation to 65th Annual Conference of the American Association for Public Opinion Research

Prepared by
Joe Baker Claudia Gentile Jason Markesich Shawn Marsh

MATHEMATICA
Policy Research, Inc.

Interviewer Monitoring: The Foundation of Data Quality

- Interviewer monitoring is used by the majority of survey organizations to evaluate interviewers' performance
- Little research has been devoted to understanding the behavior of monitors
- To explore monitors' behavior we asked monitoring staff to evaluate recordings of 8 actual telephone interviews:
  - We compared the monitors' evaluations to one another and a criterion group.
  - We conducted post-monitoring focus groups.

Research Questions

1. On what typical behavioral issues do monitors focus when evaluating interviews?
2. What factors influence monitors' ratings?
3. Are monitors coding nonstandardized interviewing behavior in an accurate and consistent manner?
4. What is the extent of between monitor variation?

Monitoring Standards and Procedures

- The monitoring system enables staff to listen to telephone interviews and view an interviewer's screen while an interview is in progress
- All telephone interviews are digitally recorded to allow for after-the-fact monitoring
- The purpose is to identify interviewer behavioral problems, such as
  - Inaccurate presentation of information about the study
  - Errors in reading questions
  - Biased probes
  - Inappropriate use of feedback when responding to questions

Monitoring Standards *(continued)*

- **Monitors evaluate interviews using an electronic monitoring form**
- **Monitoring staff provide feedback to the interviewers after each monitored interview**
- **Mathematica's standard practice is to monitor at least 10 percent of each interviewer's hours**

Monitoring Codes

- **15 separate codes indicate specific behaviors needing improvement, organized into 4 categories**
  - **Question Asking** (i.e., wording changes, skipping questions)
  - **Probing** (i.e., insufficient probing, leading, over-probing)
  - **Feedback** (i.e., inappropriate feedback, failure to provide feedback)
  - **Coding or Data Entry** (i.e., incorrect entry, coding and data entry errors)

Additional Monitoring Codes

- We used 2 codes for general positive and general nonstandard behaviors.
- Monitors also assess general voice and rapport by assigning an “S” (Standard) or “NS” (Nonstandard) to the interviewer’s
  - Volume
  - Pace
  - Clarity
  - Tone
  - Rapport

Overall Evaluation Scale

- 1 (Poor)
- 2 (Does Not Meet Expectations)
- 3 (Meets Expectations)
- 4 (Very Good)
- 5 (Excellent, “Above and Beyond”)

Participants and Procedures

- We selected 8 active monitors for the experiment.
- To establish a criterion, 3 gold standard supervisors/monitors were selected.
- Both groups followed the same procedures they would use during a typical monitoring task.
  - Monitors were instructed to regard recordings as if monitoring a live interview.
  - Monitors were instructed to fill out and submit evaluation form as they normally would.

Procedures, continued

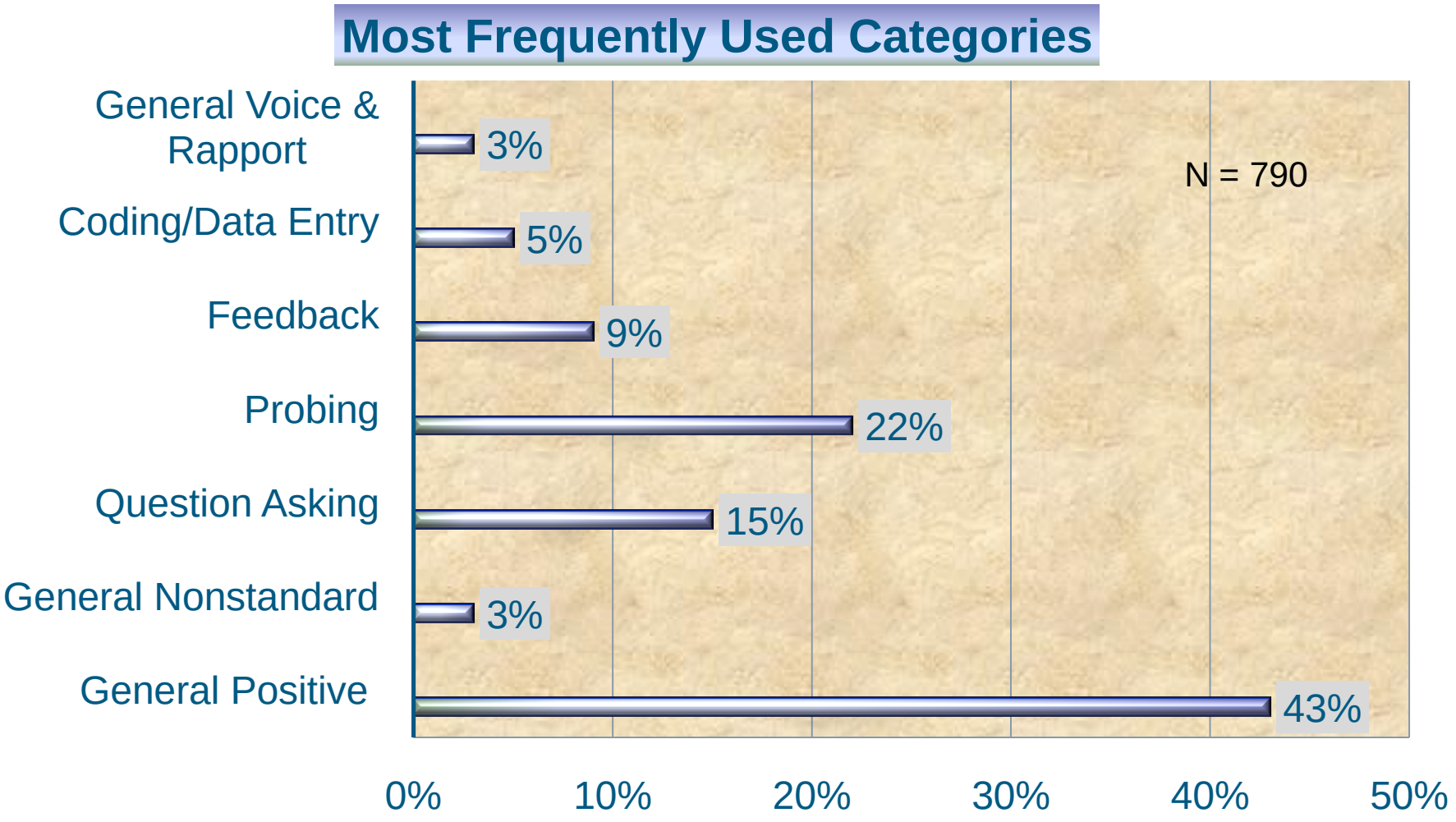
- We selected a purposeful sample of 3 complete and 5 partial interviews.
- To ensure a range of interviewing performance, we selected interviews from staff whose past performance was "above average," "average," and "below average."
- The 11 monitors were instructed to monitor the 8 recordings independently, resulting in 88 total observations, and were not permitted to discuss responses.

Summary of Recorded Interviews

Interviews Complete/Partial	Length of Time (Minutes)	Type of Interviewer
Complete	22:51	Above Average
Complete	17:21	Above Average
Partial	15:00	Average
Partial	12:00	Average
Complete	21:29	Average
Partial	15:00	Below Average
Partial	15:00	Below Average
Partial	10:44	Below Average

Note: Data is based on 8 monitors x 8 sessions = 64 monitor observations; 3 gold standard monitors x 8 sessions = 24 gold standard observations, for a total of 88 observations.

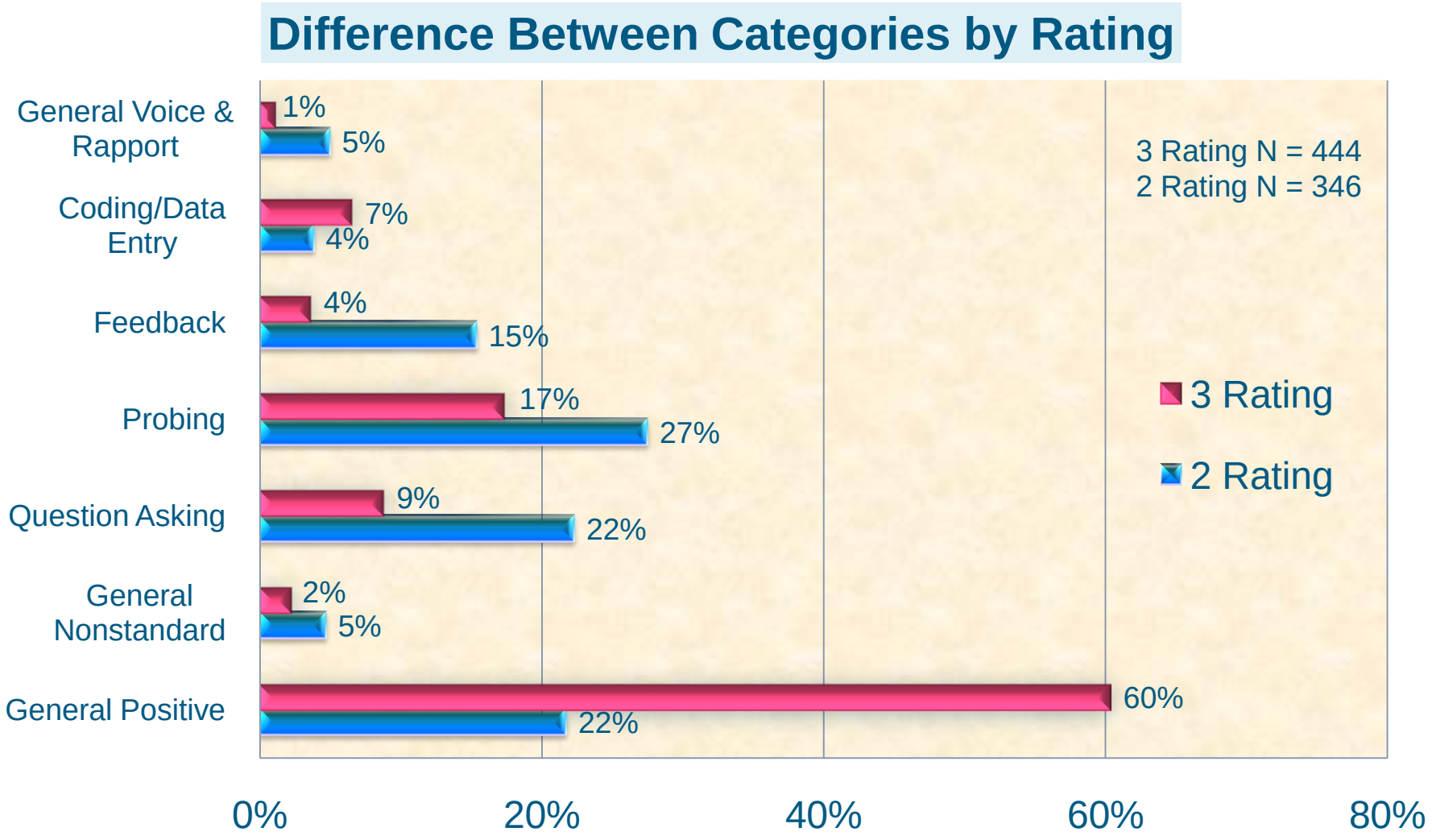
On What Typical Behavioral Issues Do Monitors Focus When Evaluating Interviews?



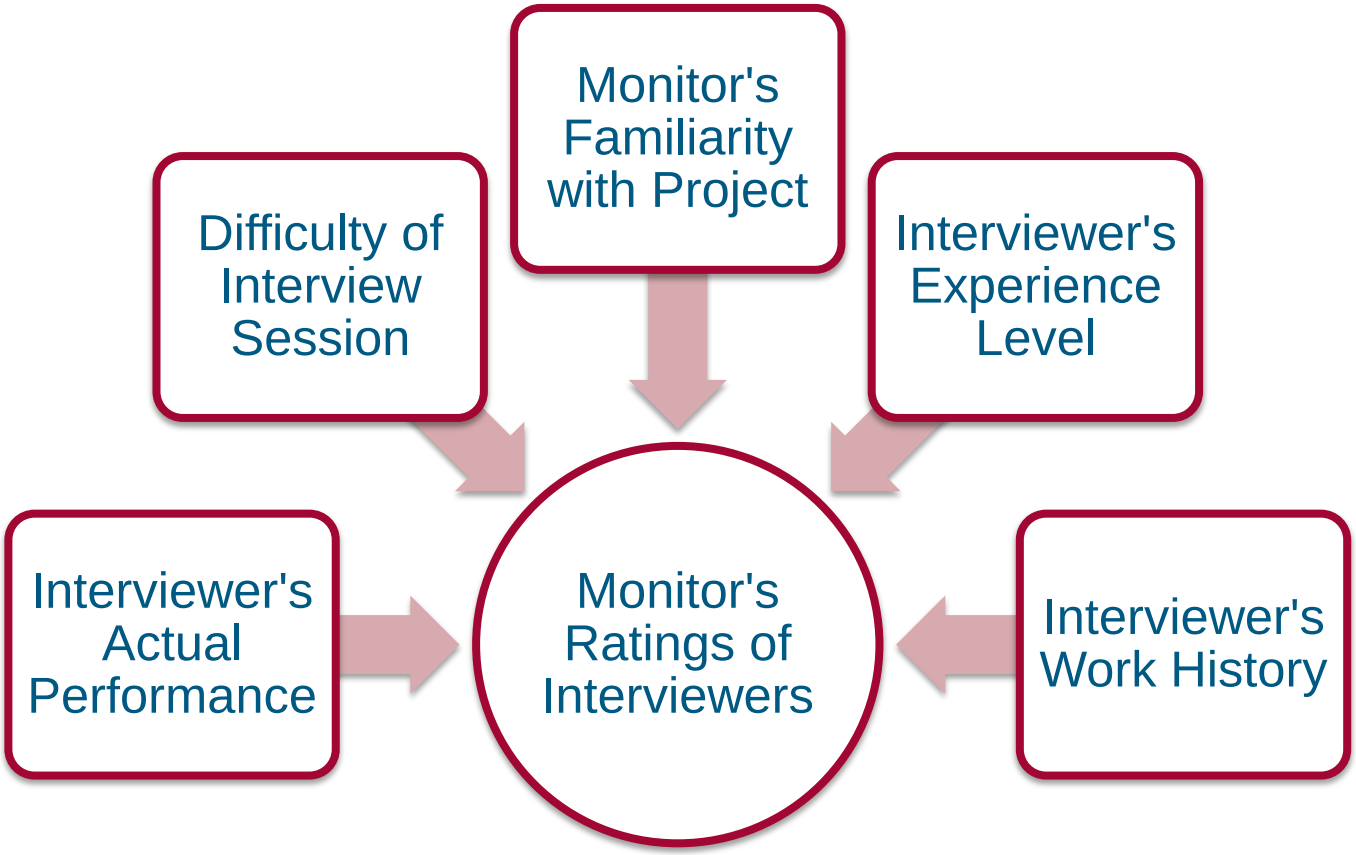
Typical Behavioral Issues that Monitors Focus on When Evaluating Interviews: Differences Between Gold Standard and Monitor Groups

Types of Codes	Gold Standard	Active Monitors	Difference
General Positive	35%	47%	-12%
Probing	26%	20%	6%
Question Asking	20%	12%	8%
Coding/Data Entry	7%	5%	2%
Feedback Error	6%	10%	-4%
General Nonstandard	4%	3%	1%
General Voice and Rapport	2%	3%	-1%

On What Typical Behavioral Issues Do Monitors Focus When Evaluating Interviews?



What Factors Influence a Monitor's Ratings?



Inter-Rater Agreement

Type of Agreement	N	Observations	Total
Agreement among gold standard monitors	3	24	96%
Agreement among active monitors	8	64	81%
Overall agreement, all monitors	11	88	89%

Note: Data is based on 8 monitors x 8 sessions = 64 monitor observations; 3 gold standard monitors x 8 sessions = 24 gold standard observations, for a total of 88 observations.

What is the Extent of Variation Between Monitors and Within Each Monitor?

- **Between Monitors**
  - **When looking at each individual evaluation form we see a clustering of the same error codes that is consistent.**
- **Within Each Monitor**
  - **It is difficult to determine in the absence of a larger sample, longer time frame, and question-by-question analysis.**

Additional Lessons From Focus Groups

- **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 and who help improve completion 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.**

Questions For Further Exploration

- If monitors use ratings of "2" and "3" differently for experienced and novice interviewers, are these ratings more a communication tool than an evaluation tool?
- If the only way to achieve a "4" is when the interview is challenging, and a "1" or a "5" is rarely assigned, is the 1–5 scale really useful?
- Would it be more helpful to monitors and interviewers if we replaced the numbered scale with more direct feedback statements? (for example, "Needs immediate attention," "needs retraining in one or two areas," "no issues, excellent job")

Thank You

- **Special thanks to the BSF, Baby FACES, and ITA 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, Walter Williams, Marianne Stevenson, and the Survey Operations Center Monitoring staff.**

- **For additional information, email Joe Baker at JBaker@mathematica-mpr.com.**