Variability in Error Detection among Telephone Monitors

Doug Currivan, Derek Stone, Kim Ault, Curry Spain and Nicole Tate

RTI International
RTI’s Quality Evaluation System (QUEST) uses a standardized protocol for live and CARI monitoring to meet multiple goals:

1. Standardization of monitoring protocols, metrics, and feedback mechanisms
2. Increased efficiency of monitoring operations
3. Increased use of CARI technology to evaluate and improve interviewer performance
4. Collection of trend data to evaluate interviewer and survey item-level performance
5. **Collection of data to evaluate variability among monitors in detecting interviewer errors**
CALL CENTER MONITORING PROCEDURES

QUEST standard procedures for call center monitoring:

1. For live sessions, monitors simultaneously view the CATI screen and listen to the interview/interaction. For recorded sessions, monitors play back recorded audio from full/partial interviews.

2. Monitors listen to live or recorded sessions for up to 12 minutes, when the session involves an interview in progress. (The first two recorded sessions for each interviewer that involve an interview are listened to in their entirety.)

3. Monitors enter any interviewer errors observed under the appropriate interviewing skill area.

4. QUEST automatically produces overall and skill area scores based on the number and severity of errors coded by the monitor.

5. Monitors deliver immediate feedback to interviewers on their overall session score and skill area scores.
How Much and When Might Monitors Vary?

Our assumption is that standardization should tend to produce consistent ratings among telephone monitors (Couper, et al. 1992; Fowler & Mangione, 1990)

1. To what extent do monitors vary in observing interviewer errors over the course of a field period? Do any monitors appear to be notably “hard” or “easy” raters?

2. To what extent do monitors vary in detecting interviewer errors for the skill areas where errors are most common?

3. To what extent do monitors agree on the number and type of errors committed when rating the same set of interviewing sessions?
1. Overall Monitor Variability in Error Detection

Reports on variability in error detection among a set of monitors on a single study

- for all monitors and for the entire field period
- by monitor experience level for the entire field period

Key assumptions and considerations:
- same instrument/protocol throughout the field period
- random assignment of sessions to monitors
- minimum number of sessions completed set to 25+
- sessions involved a complete or partial interview
Variability in Monitors’ Overall Mean Error Detection Rates

Mean Proportion of Sessions with at Least One Error Detected

Number of Monitoring Sessions Completed

total monitors = 23, total sessions = 2,031

mean = 0.506, sd = 0.111
Variability in Overall Error Detection Rates by Experience Level

<table>
<thead>
<tr>
<th>Experience Level</th>
<th>Monitors</th>
<th>Sessions</th>
<th>Mean Proportion</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 2 Years</td>
<td>11</td>
<td>1,485</td>
<td>0.506</td>
<td>0.104</td>
</tr>
<tr>
<td>2 or More Years</td>
<td>12</td>
<td>546</td>
<td>0.506</td>
<td>0.132</td>
</tr>
</tbody>
</table>

Mean Proportion of Sessions with at Least One Error Detected

Number of Monitoring Sessions Completed
Variability in Overall Error Detection Rates by Experience Level

**Less than 2 Years Experience**

- **monitors = 11, sessions = 1,485**
  - mean = 0.506, sd = 0.104

**2 or More Years Experience**

- **monitors = 12, sessions = 546**
  - mean = 0.506, sd = 0.132

Mean Proportion of Sessions with at Least One Error Detected

Number of Monitoring Sessions Completed
2. Variability within Interviewing Skill Areas

Error detection variability for the four interviewing skill areas with the highest interviewer error rates:

- **probing skills** – leading/non-neutral probes, insufficient probing, not probing at all, etc.
- **questionnaire administration** – not using correct interviewing techniques, omitting questions, etc.
- **interview protocol** – insufficient study knowledge, required study materials not used, etc.
- **initial contact** – inadequate responses to questions, refusal aversion techniques not used, etc.
Variability in Mean Error Detection Rates: Probing Skills

- Total monitors = 23
- Total sessions = 2,031

Mean Proportion of Sessions with at Least One Probing Skills Error

- Mean = 0.281
- Standard deviation = 0.095

Number of Monitoring Sessions Completed

- Graph shows a line with points scattered around it, indicating variability in error detection rates.

RTI International
Variability in Mean Error Detection Rates: Questionnaire Admin.

- **total monitors = 23, total sessions = 2,031**
- **mean = 0.265, sd = 0.123**

Number of Monitoring Sessions Completed vs. Mean Proportion of Sessions with at Least One Questionnaire Administration Error.
Variability in Mean Error Detection Rates: Questionnaire Admin.

- Total monitors = 23, total sessions = 2,031
- Mean proportion of sessions with at least one questionnaire administration error:
  - Mean = 0.265
  - Standard deviation (sd) = 0.123
3. Test of Monitor Agreement on Error Detection

Seeded the same 10 abbreviated interview sessions into the workloads of 11 monitors to see to what extent and in what skill areas they agreed on errors committed.

1. examined range of agreement coefficient across all 10 interviewing skill areas
2. highlighted agreement rates for the 4 interviewing skill areas with highest interviewer error rates
3. used SAS macro from Blood & Spratt (2007) to calculate Gwet’s AC2 (2001) for multiple raters and multiple ordinal categories
Monitors’ error detection produces skill area scores based on both the number and severity of interviewer errors detected by monitors:

1 = no errors

2 = some errors (none critical)

3 = excessive errors (multiple errors and/or 1 or more critical errors)
## AC2 Values for All Interviewing Skill Areas

<table>
<thead>
<tr>
<th>Interviewing Skill Areas</th>
<th>AC2 Value</th>
<th>Agreement Level*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authenticity</td>
<td>0.899</td>
<td>Excellent</td>
</tr>
<tr>
<td>Case Management</td>
<td>0.899</td>
<td>Excellent</td>
</tr>
<tr>
<td>Initial Contact</td>
<td>0.437</td>
<td>Good</td>
</tr>
<tr>
<td>Keying Skills</td>
<td>0.864</td>
<td>Excellent</td>
</tr>
<tr>
<td>Questionnaire Administration</td>
<td>0.115</td>
<td>Poor</td>
</tr>
<tr>
<td>Probing Skills</td>
<td>0.208</td>
<td>Poor</td>
</tr>
<tr>
<td>Feedback Skills</td>
<td>0.817</td>
<td>Excellent</td>
</tr>
<tr>
<td>Presentation Skills</td>
<td>0.817</td>
<td>Excellent</td>
</tr>
<tr>
<td>Professional Behavior</td>
<td>0.640</td>
<td>Good</td>
</tr>
<tr>
<td>Interview Protocol</td>
<td>0.322</td>
<td>Poor</td>
</tr>
</tbody>
</table>

* Fleiss (1981) guidelines for reliability coefficients: < 0.40 = poor, 0.40-0.75 = good, > 0.75 = excellent
Some Conclusions

1. Data on overall mean error detection rates and number of sessions completed highlighted some differences between experienced and inexperienced monitors
   - experienced monitors had somewhat greater variability in mean overall error detection rates, but also completed fewer sessions on average

2. Data on mean error detection rates for interviewing skill areas with the highest error rates identified only 1 monitor who appeared to be a notably “hard” rater
   - 1 monitor had the highest error detection rates for any errors and for one skill area (questionnaire administration)
More Conclusions

3. Results from a blind test of monitor agreement on number and types of errors committed in 10 interviews showed mixed agreement rates across interviewing skill areas.

4. In general, monitor agreement appeared higher for more routine interviewing tasks and lower for higher-level interviewing skills.

5. The results of the agreement test indicated greater variability for higher-level interviewing skills than was observed in reports of monitor variability in the field.
Further Research

1. Continue to examine monitor variability in error detection for other surveys with different protocols, question types, etc.

2. Conduct further tests of monitor agreement
   - compare monitors to “gold standard” session results
   - reconcile monitor disagreements with each other and differences from gold standard

3. Conduct multivariate analysis to determine what factors are most strongly associated with variation in monitors’ detection of errors.
Acknowledgements

Monitoring team members:
   Susan Kinsey, Howard Speizer, Richard Heman-Ackah,
   Sridevi Sattaluri, Dave Foster, and Melissa Cominole

Advice on calculating agreement coefficients:
   Carla Bann and Jamie Newman
Presenter Contact Information

Doug Currivan
dcurrivan@rti.org
919.316.3334