2013 American Association for Public Opinion Annual Conference

Variability in Error Detection among Telephone Monitors

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

Abstract

Standardized methods and tools for monitoring telephone interviewers are important for ensuring survey data meet high quality standards. In order to effectively limit the risk of interviewer behaviors biasing or adding variance to survey estimates, the quality monitoring process requires accurate and consistent detection of interviewer errors. To this end, RTI has developed a standardized, mode-independent interview quality monitoring evaluation system, QUEST. This system supports evaluation of interviewing quality through both live monitoring and review of digitally-recorded sessions. QUEST allows telephone interviewing behaviors to be evaluated using a common set of quality metrics that are stored in a single shared database. These metrics are based on objective indicators of specific interviewer behaviors, including definitions and concrete examples for each behavior, as opposed to more subjective ratings or impressions of interviewing quality.

The primary hypothesis of our research is that the standardized, objective approach followed in QUEST will produce minimal variation across monitors in their detection of interviewer errors and other unacceptable behaviors. Two primary sources of data are used to investigate variability in the rates at which monitors detect interviewer errors: comparison of error detection rates across monitors from monthly monitoring results and examination of the results of blind scoring by monitors of a set of 10 selected interviewing scenarios. Comparisons of error detection rates across monitors includes both overall errors detected across sessions and errors detected for specific interviewing skill areas. In addition, this analysis examines whether scoring across monitors varies when factors such as interviewing shifts or monitor experience levels are considered. Based on the results of the comparisons of monthly monitor scores and blind scoring of interviewing scenarios, this presentation discusses the implications of the observed levels of monitor scoring variability in general and disagreements on specific scenarios for accurate and consistent detection of interviewing errors.

1. Background

Nearly twenty years ago, Couper, Holland, and Groves (1992) noted that monitoring protocols often (1) followed unsystematic and subjective procedures and (2) included only general impressions of telephone interactions, rather than objective measures of behavior. In recent years, standardizing methods and tools for evaluating the quality of survey interviewing across modes and studies has increasingly been an important goal for survey organizations. RTI has developed a standardized, mode-independent interview quality monitoring evaluation system, QUEST (Currivan, et al. 2011, Speizer, et al. 2009; Speizer, et al. 2010). This system allows in-person and telephone interviewing behaviors to be evaluated using a common set of quality metrics that are stored in a single shared database. The system supports evaluation of interviewing quality for both live monitoring in real time and review of computer audio-recorded interview (CARI) files.

QUEST monitoring of interviews in RTI's call center follows a set of standard procedures:

- 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.
- 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.)
- Monitors enter any interviewer errors observed under the appropriate interviewing skill area.
- QUEST automatically produces overall and skill area scores based on the number and severity of errors coded by the monitor.
- Monitors deliver immediate feedback to interviewers on their overall session score and skill area scores.

QUEST was designed to meet multiple goals to support interviewing quality:

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 (Biemer, et al. 2000, Thissen, et al. 2008)
4. Collection of trend data to evaluate interviewer and survey item-level performance (Couper, et al. 1992, Hicks, et al. 2010).
5. Collection of data to evaluate variability among monitors in detecting interviewer errors

This paper focuses on the fifth objective by examining telephone monitor variability for an entire field period and analyzing the results of a blind test of monitor agreement on 10 interviewing scenarios. Our assumption is that standardization of quality monitoring under QUEST should tend to produce consistent ratings among telephone monitors (Couper, et al. 1992; Fowler & Mangione, 1990). This paper examines the following three research questions on the variability of monitors in detecting interviewing errors:

1. To what extent do monitors vary in observing interviewer errors over the course of a field period? More specifically, do any monitors appear to be notably "hard" or "easy" raters to the extent that their ratings appear biased?
2. To what extent to 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?

Based on monitoring results from an entire field period of a survey, ***Section 2*** of this paper addresses monitor variability in detecting any interviewer errors, differences in error detection based on monitor experience, and detecting errors within specific interviewing skill areas. ***Section 3*** examines the results of a blind test of monitor agreement for all interviewing skill areas. ***Section 4*** summarizes the key findings from this research and suggests further research for continued assessment of monitor variability in detecting interviewer errors.

2. Evaluation of Variability among Monitors in Detecting Errors

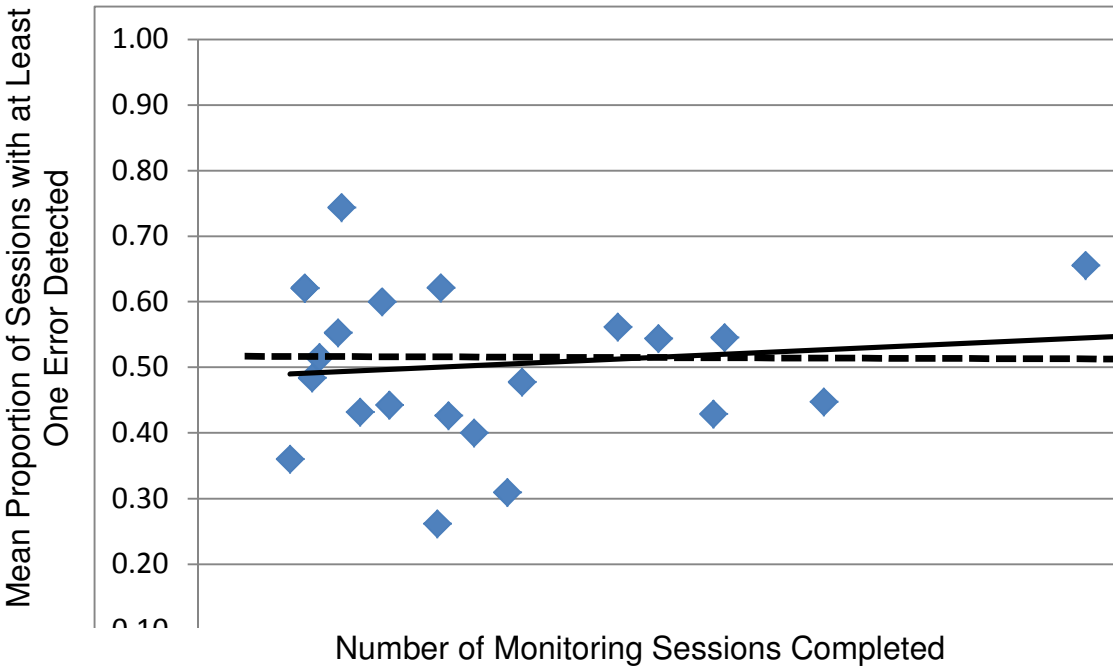
Our initial investigation of monitor variability in error detection examined the distribution of monitors' error detection rates for the entire field period of a telephone survey and also considered factors such as monitor experience level and weekday versus evening and weekend shifts. A few key assumptions and considerations guided this investigation:

- the same survey instrument and protocol was maintained throughout the field period,
- monitoring sessions were assigned randomly to monitors,
- all sessions included in the analysis involved a either a complete or partial interview, and
- all monitors included in the analysis completed at least 25 sessions during the survey.

Our initial assessment of inter-rater reliability used QUEST reports to examine patterns in monitors' mean error detection rates for any error and for specific interviewing skill areas. ***Exhibit 1*** shows the pattern of mean error detection rates for any interviewing errors for the complete set of 23 monitors who completed at least 25 monitoring sessions (live or recorded) during the entire field period. The number of monitoring sessions conducted by these 23 monitors ranged from 29 to 282 sessions.

The mean error detection rates ranged from 0.262 of monitoring sessions completed to 0.744 of sessions. As the scatter plot indicates, most monitors were relatively close to the overall mean error detection rate of 0.506. The standard deviation for the overall mean error detection rate was 0.111, giving a coefficient of variation of 21.9. The trend line for monitors' mean error detection rates by the number of monitoring sessions shown in ***Exhibit 1*** indicates no clear association between mean detection rates and number of monitoring sessions. An association between mean detection rates and number of monitoring sessions was not anticipated, but these plots were created to examine whether such a correlation might exist and, therefore, potentially inform the interpretation of these results. The trend line had a slight upward slope, which appears unlikely to represent a meaningful association. Moreover, with only 23 monitors, the pattern is skewed somewhat by the monitor with the highest mean error detection rate (0.744) and the two monitors with the highest numbers of sessions (241 and 282).

Exhibit 1
Variability in Monitors' Overall Mean Error Detection Rates for the Entire Field Period

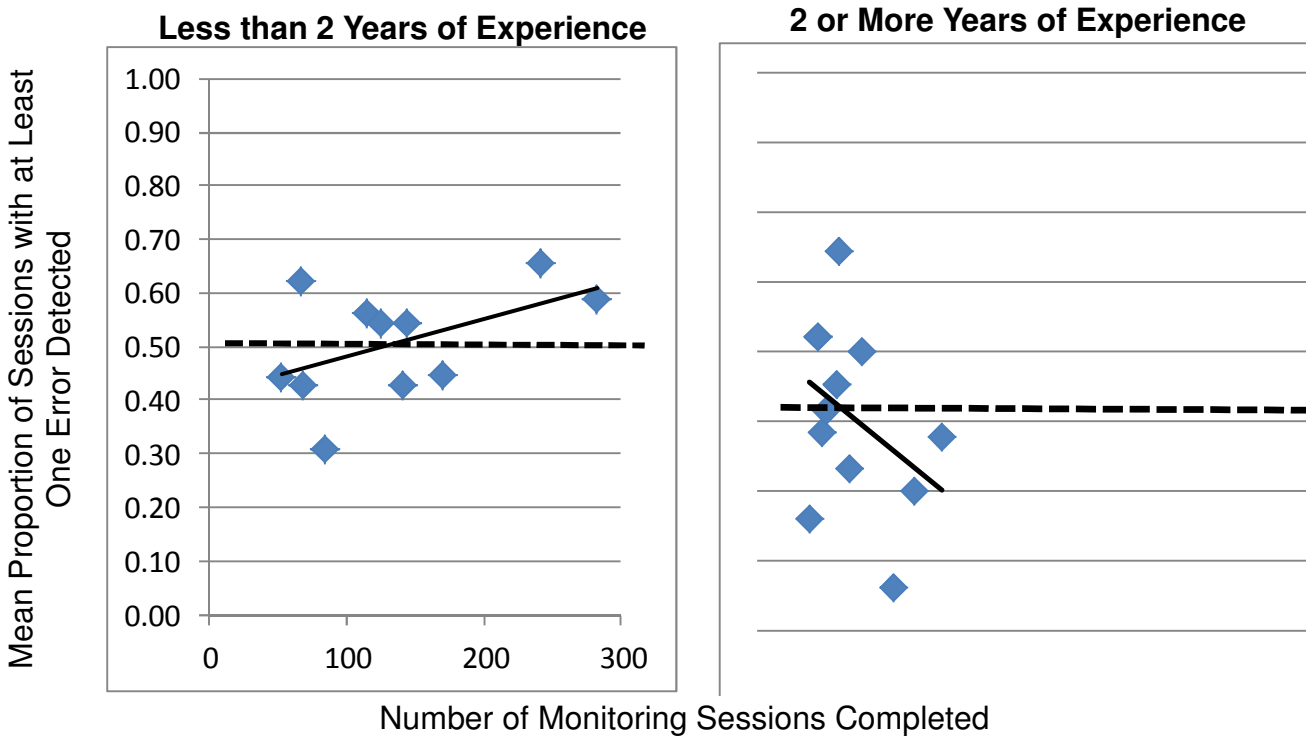


QUEST reports also allow for examining monitors' error detection rates by (1) monitor and experience level and (2) weekday versus evening and weekend interviewing shifts. Because over two-thirds of the 23 monitors who completed at least 25 monitoring sessions during the entire field period worked evening and weekend shifts, this comparison was not very balanced. Among these 23 monitors, a nearly equal set had less than two years of monitoring experience (11) versus those who had two or more years of experience (12).

Exhibit 2 presents scatter plots of mean error detection rate by number of monitoring sessions separately for monitors with less than two years of experience and those with two or more years of experience. These plots reveal a number of important differences between inexperienced and experienced monitors:

1. The mean detection rate for any error was identical (0.506), but the standard deviation for experienced monitors (0.132) was greater than the standard deviation for inexperienced monitors (0.104). The coefficient of variation was therefore somewhat greater for experienced monitors (26.1) than for inexperienced monitors (20.6).
2. On average, inexperienced monitors completed a much higher number of monitoring sessions (135) than experienced monitors (45.5). This contrast was exacerbated by the two inexperienced monitors who conducted a much higher numbers of sessions than any other monitors.
3. The trend line for inexperienced monitors suggest a slight increase in error detection rates occurred as the number of monitoring sessions increased. For experienced interviewers the short trend line suggests the opposite, a significant decrease in error detection rates as the number of monitoring sessions increased. It should be noted that the range of monitoring sessions completed by experienced monitors was much more limited, from 29 to 88.
4. The experienced monitor with the highest mean error detection rate (0.744) appears to have a significantly higher rate than any of the other experienced or inexperienced monitors.

Exhibit 2
Variability in Monitors' Overall Mean Error Detection Rates by Monitor Experience Level



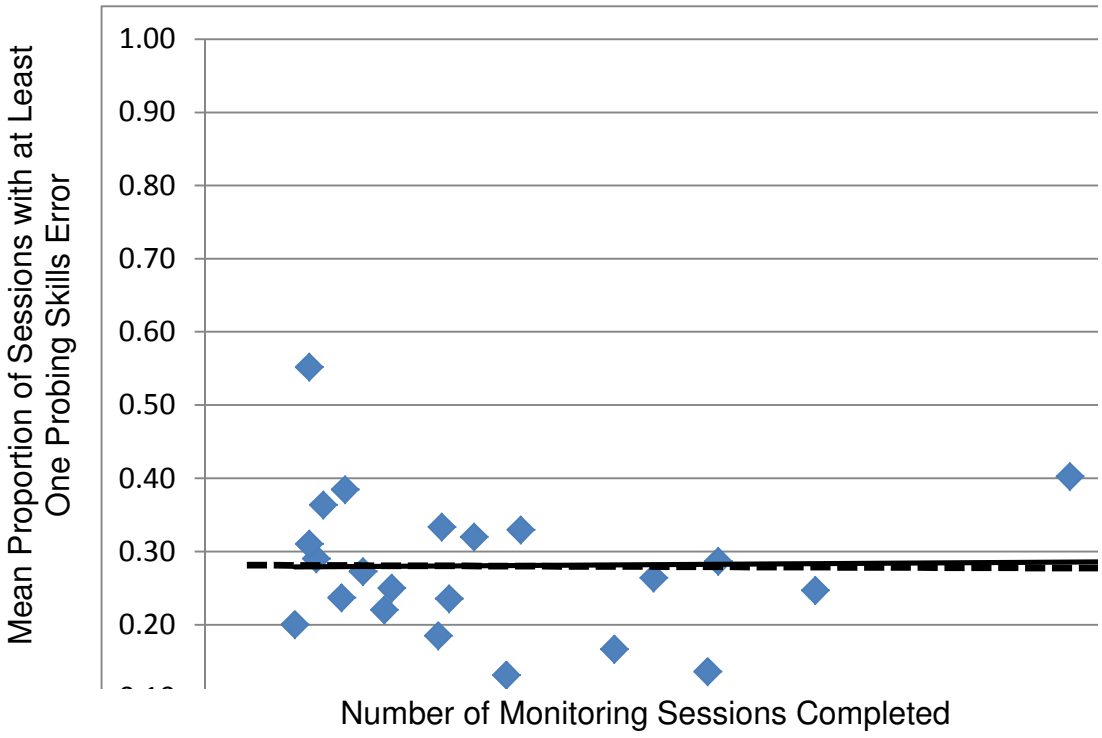
In addition to overall interviewer error detection, examining how monitors vary with respect to detect errors within specific skill areas was a second important objective of this research. The focus was on variability in monitors' error detection rates for the interviewing skill areas that had the highest rates of at least one error in the skill area being detected. Across the entire field period, the following four interviewing skill areas had the highest mean error detection rates:

1. Probing skills (0.281) – included seven QUEST items on leading or non-neutral probes, insufficient probing, not probing at all, and other probing issues.
2. Questionnaire administration (0.265) – included six QUEST items on not using correct interviewing techniques, omitting questions, and other administration problems.
3. Interview protocol (0.134) – included nine QUEST items on insufficient study knowledge, required study materials not used, and further protocol elements.
4. Initial contact (0.112) – included seven QUEST items on inadequate responses to questions, refusal aversion techniques not used, and other problems during initial contact with sample members.

Given that about 79 percent of all monitoring sessions included detection of at least one error in one or more of these skill areas, the impact of monitor variability in these skill areas had a substantial potential to influence the overall monitoring results.

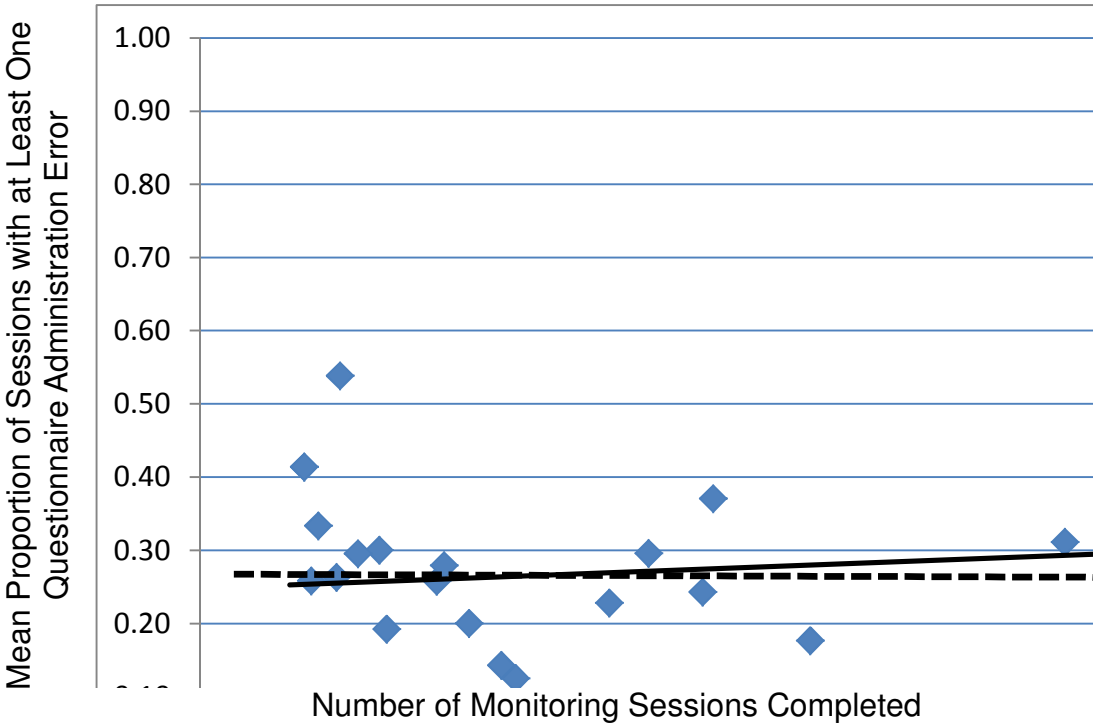
Exhibit 3 presents a scatter plot showing the mean error detection rate for Probing Skills by the number of monitoring sessions completed during the entire field period. With the exception of one monitor with a mean error detection rate of 0.552, most of the monitors' rates clustered fairly close to the mean of 0.281. The standard deviation was 0.095, resulting in coefficient of variation equal to 33.3. The trend line was virtually horizontal, indicating no association between the mean error detection rate and the number of monitoring sessions completed by the 23 monitors.

Exhibit 3
Variability in Monitors' Mean Error Detection Rates: Probing Skills



The scatter plot in ***Exhibit 4*** shows the mean error detection rate for Questionnaire Administration by the number of monitoring sessions completed during the entire field period. In contrast with Probing Skills, the range of mean error detection rates varied from 0.000 to 0.538. Although most of the monitors' rates clustered fairly close to the mean of 0.265, monitor variability for Questionnaire Administration was greater than Probing Skills. The standard deviation was 0.123, resulting in coefficient of variation equal to 46.4. Similar to the overall error detection plot in ***Exhibit 1***, the trend line had a slight upward slope that appears unlikely to represent a meaningful association. One further observation from ***Exhibit 4*** is that the monitor who had the highest detection rate (0.538) for Questionnaire Administration was the same monitor who had the highest overall error detection rate (0.744) in ***Exhibit 1***. This monitor was the only one identified as arguably having an (upward) bias in her or his interviewer error detection rates. Given that this monitor had error detection rates close to the mean for other interviewing skill areas, evidence for an upward bias across all of this monitors' work was not conclusive.

Exhibit 4
Variability in Monitors' Mean Error Detection Rates: Questionnaire Administration



3. Testing Monitor Agreement on Error Detection for All Interviewing Skill Areas

In September 2012, call center staff seeded the same 10 recordings of abbreviated interview sessions into the regular QUEST workloads of the 11 monitors who were working on the survey at that time. During the last month of the survey, researchers and interviewers offered sample members who had yet to complete the full interview the option to complete an abbreviated interview, as a tool to encourage participation in the final weeks of the field period. The abbreviated interviews provided a convenient opportunity to use these short interviews for the test of monitor agreement. The 10 abbreviated interviews were selected purposefully, with the goal of including sessions with multiple interviewer errors and a variety of error types. Interviewer and respondent characteristics were not considered in selecting sessions. These sessions appeared exactly like the other sessions included in the monitors' QUEST

workloads, so they had no indication that these sessions were unique. Thee monitors were unaware that these same set of sessions were assigned to 10 other monitors.

The primary the goal of the test of monitor agreement was to obtain a sense of how a team of monitors agreed on interviewer errors committed within each skill area. For this analysis, we examined the range of agreement coefficient across all 10 interviewing skill areas. Furthermore, we again focused special attention on the four interviewing skill areas that had the highest error detection rates over the course of the field period – probing skills, questionnaire administration, interview protocol, and initial contact

The literature on inter-rater reliability provides several different statistics for calculating agreement levels among a set of raters. Cohen's (1960) Kappa and weighted Kappa (1968) are common statistics used to measure inter-rater reliability, but Kappa suffers from multiple limitations, as noted by Uebersax (2002);

1. Kappa is not a chance-corrected measure of agreement.
2. Kappa is an omnibus index of agreement that does not distinguish among various types or sources of disagreement.
3. Kappa is influenced by trait distribution (prevalence) and base rates, making Kappas incomparable across studies.
4. Kappa may be low even though there are high levels of agreement and individual ratings are accurate.
5. With ordered category data, weights must be selected arbitrarily to calculate weighted Kappa.

These considerations were directly relevant to calculating agreement coefficients for the blind test of monitor agreement. This test involved multiple raters who produced one of three ordinal scores for each interviewing skill area in QUEST:

1 = no errors observed

2 = some errors observed (none critical)

3 = excessive errors observed (multiple errors and/or one or more critical errors)

In addition, an important goal was to assess agreement levels for each interviewing skill area and be able to reliably compare agreement rates across skill areas. Although concerns about chance correction were relatively low given the nature of the monitoring process, correcting for chance agreement avoids inappropriately over-stating agreement levels.

Considering these factors, second-order agreement coefficient (AC2 – Gwet, 2001) statistics were calculated, using the AC1AC2 SAS macro (Blood & Spratt, 2007). The AC2 agreement coefficient addresses concerns about Cohen's Kappa statistic, especially obtaining a low Kappa when inter-rater agreement is high and correctly calculating agreement coefficients for ordinal scales. To interpret the meaning of the AC2 agreement coefficient, we followed the general guidelines for reliability coefficients recommend by Fleiss (1981):

< 0.40 = poor agreement

0.40 - 0.75 = good agreement

> 0.75 = excellent agreement

Based on the AC2 values presented in ***Exhibit 5***, the interviewer skills that had excellent agreement levels include authenticity, case management, feedback skills, keying skills, and presentation skills. Inter-rater agreement was good for initial contact (0.437) and professional behavior (0.640) skills. Agreement levels

were poor for the remaining three skill areas – interview protocol, probing skills, and question administration. These skill areas represent three of the four areas that had the highest error detection rates over the course of the field period. Poor monitor agreement for most of the skill areas with high error detection rates suggests that future tests of agreement should continue to focus on these areas. These areas involve higher-level interviewer skills that might invoke significant disagreement among monitors on appropriate interviewing practices.

Exhibit 5
Second Order Agreement Coefficients (AC2) for All Interviewing Skill Areas

Interviewing Skill Areas	AC2 Value	Agreement Level*
Authenticity	0.899	**Excellent**
Case Management	0.899	**Excellent**
Initial Contact	0.437	***Good***
Keying Skills	0.864	**Excellent**
Questionnaire Administration	0.115	Poor
Probing Skills	0.208	Poor
Feedback Skills	0.817	**Excellent**
Presentation Skills	0.817	**Excellent**
Professional Behavior	0.640	***Good***
Interview Protocol	0.322	Poor

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

4. Some Conclusions and Next Steps

Examination of patterns of monitor variability in error detection rates and the results of a blind test of monitor agreement suggested the following 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. To the extent that a non-trivial correlation existed between error detection rates and number of sessions completed, this association confound interpretation of the comparison of error detection rates between inexperienced and experienced monitors.
2. Data on mean error detection rates for interviewing skill areas with the highest error rates identified only one monitor who could potentially be viewed as having an upward bias in detecting interviewer errors. This monitor had the highest error detection rates for any errors and for one skill area (questionnaire administration), but had error detection rates close to the mean for other interviewing skill areas. As a result, evidence for an upward bias across all of this monitors' work was not conclusive.
3. The results from a blind test of monitor agreement on number and types of errors committed in 10 abbreviated interviews showed mixed agreement rates across the 10 interviewing skill areas. In general, monitor agreement appeared higher for more routine interviewing tasks and lower for higher-level interviewing skills. The results of the agreement test suggested greater variability for higher-level interviewing skills than was observed from QUEST reports on these skill areas for the survey field period.

These conclusions suggest further research needed to improve our understanding of monitors' variability on interviewer error detection and, therefore, survey data quality:

1. Continue to examine monitor variability in error detection for other surveys with different protocols and question types. Existing QUEST reports can be used to provide a sense of when and how monitors' vary, and what impact this might have on providing accurate an effective feedback to interviewers to minimize errors over the course of a field period.
2. Conduct further tests of monitor agreement for other surveys with different sets of monitors. Two further steps that will enhance tests of monitor agreement are (1) to compare monitors to "gold standard" session results determined by an expert panel and (2) reconcile monitor disagreements with each other and differences from the gold standard through group discussions. These further steps are likely to yield useful information for training and supervising monitor teams.
3. Conduct multivariate analysis to determine what factors are most strongly associated with variation in monitors' detection of errors. Recent research by Baker, et al. (2013) indicates that monitors include a wide range of factors in rating interviewers' work, not just technical criteria. Using monitoring sessions as the unit of analysis, modeling the factors most likely to predict error detection could answer some key questions on monitor behavior even within the highly standardized protocols created by QUEST. For example, does interviewer experience level have any impact on the probability of monitors detecting specific errors? Or do monitor characteristics have a greater impact on when and where they detect errors?

Given that research using the QUEST monitoring database is still at an early stage, continued efforts analyze monitor data seem likely to improve our understanding of monitor behavior. Further analysis seems promising for informing monitor training and supervision in pursuit of maximizing survey data quality.

5. Acknowledgements

The authors thank other current members of the QUEST team: Susan Kinsey (Lead), Howard Speizer, Richard Heman-Ackah, Sridevi Sattaluri, Dave Foster, and Melissa Cominole. We also gratefully acknowledge the advice of Carla Bann and Jamie Newman on calculating agreement coefficients.

6. References

Baker, J., Gentile, C., Markesich, J., Marsh, S., Panzarella, E., and Weiner, R. (2013). Ensuring data quality: What criteria do monitors use to rate interviewers? *Survey Practice*, vol. 6, no. 1.

Biemer, P., Herget, D., Morton, J., and Willis, W.G. (2000). The feasibility of monitoring field interview performance using computer audio recorded interviewing (CARI). In *Proceedings of the American Statistical Association's Section on Survey Research Methods*, pp. 1068-1073.

Blood E. and Spratt, K. (2007). Disagreement on agreement: Two alternative agreement coefficients. SAS Global Forum, Paper 186–2007. Downloaded at: http://www2.sas.com/proceedings/forum2007/186-2007.pdf .

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213-220.

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

Currivan, D., Stone, D., Fuller, K., Kinsey, S. and Speizer, H. (2011). Some Implications of Standardizing Methods for Quality Monitoring of Survey Interviewing. *Proceedings of Statistics Canada Symposium: Strategies for Standardization of Methods and Tools – How to get there*. Ottawa, ON.

Fleiss, J. (1981). *Statistical methods for rates and proportions* (Second edition). NewYork: Wiley.

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

Gwet, K. (2001). *Handbook of inter-rater reliability: How to measure the level of agreement between two or multiple raters*. Gaithersburg, MD: Stataxis.

Hicks, W., Edwards, B., Tourangeau, K., McBride, B., Harris-Kotejin, L., and Moss, A. (2010). Using CARI Tools to Understand Measurement Error. *Public Opinion Quarterly, 74*, 985-1003.

Speizer, H., Kinsey, S., Heman-Ackah, R., and Thissen, R. (2009). Developing a common, mode-independent, approach for evaluating interview quality and interviewer performance. Presented at *Federal Committee on Statistical Methodology* Research Conference, Washington, D.C.

Speizer, H., Currivan, D., Heman-Ackah, R., and Kinsey, S. (2010). Developing a common, mode-independent, approach for evaluating interview quality and interviewer performance: lessons learned. Presented at the *American Association for Public Opinion Research* Annual Conference, Chicago, IL.

Thissen, M.R., Sattaluri, S., McFarlane, E., and Biemer, P. (2008). The evolution of audio recording in field surveys.'' *Survey Practice*. http://surveypractice.org/2008/12/19/audio-recording.htm .

Uebersax, J. (2002). Kappa coefficients: A critical appraisal. http://ourworld.compuserve.com/homepages/jsuebersax/kappa.htm