What Factors Explain Variation in Monitors' Detection of Interviewing Errors in Telephone Surveys?

Douglas B. Currivan[1], Paul P. Biemer[1], Tamara Terry[1], Ivan Carillo-Garcia[1]
[1]RTI International, 3040 E. Cornwallis Road, Research Triangle Park, NC 27709

Abstract
To limit the potential for interviewer behavior to bias or add variance to survey estimates, monitoring telephone surveys requires accurate and consistent detection of interviewing errors, interview protocol violations and other biasing behaviors. Multiple factors can affect telephone monitors' detection of interviewer deficiencies. Because monitoring results alone are insufficient to understand how multiple factors contribute to monitors' error detection, this research combines data from (1) monitoring sessions, (2) responses to a survey of monitors, and (3) administrative records with monitor and interviewer characteristics. These data operationalize multiple sources of variation in monitors' error detection that could not be captured by the monitoring system data alone. Multi-level models are used to analyze the contribution of these sources to variation in monitors' detection of interviewing errors. This paper discusses the implications of these results for understanding sources of variation in monitors' detection of interviewing errors and for guiding decisions on designing monitoring processes in centralized telephone survey centers.

Key Words: telephone survey monitoring, interviewing errors, variation in interviewing error detection

1. Background and Objectives

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; Currivan, et al. 2013; 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; Biemer, et al. 2001; Thissen, 2014; Thissen, et al.2008; Thissen, et al., 2013) and other digitally-recorded files.

QUEST was designed to meet multiple goals in maintaining high data quality:

1. Standardization of monitoring and interviewing protocols, metrics, and feedback mechanisms

2. Increased efficiency of monitoring and interviewing operations

3. Increased use of CARI and other digital recordings to evaluate and improve interviewer performance (Biemer, et al. 2001; Thissen, 2014; Thissen, et al. 2008; Thissen, et al. 2013)

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 (Currivan, et al. 2011; Currivan, et al. 2013)

This paper focuses on the fifth objective. The goal of this research is to better understand what factors account for variation in monitors' detection of interviewers' protocol violations. Understanding these factors will help identify any further steps needed to improve the monitoring system, and specifically monitors' procedures for reviewing interviewers' work. Previous research on monitor variability in detecting errors using QUEST has produced the following findings:

- Two examinations of mean error detection rates for interviewing skill areas with the highest error rates generally identified only one or two monitors who appeared to detect interviewer errors at a significantly *higher* rate than all other monitors. These examinations did not identify any monitors who appeared to detect interviewer errors at a significantly lower rate than other monitors (Currivan, et al. 2011; Currivan, et al. 2013).

- Experienced monitors had somewhat greater variability in mean overall error detection rates, although the more experienced monitors had completed fewer sessions on average than less experienced monitors in the data examined. With a higher average number of sessions, variation among more experienced monitors could have possibly been reduced to levels more similar to less experienced monitors (Currivan, et al. 2013). If higher workloads for more experienced monitors did not result in more similar variation compared with less experienced monitors, this could suggest that a number of more experienced monitors were not following the standardized monitoring procedures as consistently as less experienced monitors.

- Results from a blind test of monitor agreement on numbers and types of errors committed in 10 abbreviated interviews showed mixed agreement rates across 10 interviewing skill areas. In general, monitor agreement was higher for more routine interviewing tasks such as case management, keying skills, feedback skills, and presentation skills, Monitor agreement was much lower for higher-level interviewing skills such as questionnaire administration and probing skills (Currivan, et al. 2013).

This previous research raises a number of questions about why monitors might vary in identifying interviewer protocol violations. Holding interviewer variation constant, monitor variability should be minimal under the highly standardized QUEST training, procedures, and tools. To more closely examine monitor variability, this study addresses two primary research questions:

1. How much variation in errors observed across skill areas can be attributed to monitors, as opposed to other sources such as interviews and sessions?

2. What other factors might be significantly associated with monitors' variation in detecting interviewer protocol violations across skill areas, such as live versus recorded sessions, interviewer characteristics, monitor characteristics, or monitor orientations toward their work?

Recent research by Baker, et al. (2013) indicates that monitors can include a wide range of factors in rating interviewers' work, not just the technical criteria provided to them via training and supervision. Multiple factors can affect monitors' detection of interviewing errors, including monitors' characteristics and attitudes, how interviewers are supervised, respondent characteristics and behavior, interviewer characteristics, the survey questions and protocol being monitored, the telephone survey center environment, and the monitoring system being used. QUEST data alone are insufficient to fully explain variation in monitors' detections of interviewing errors. For this reason, this research used the following data to examine what factors account for variation in monitors' detection of errors:

- QUEST monitoring session results,
- responses to a survey of monitors orientation to their work, and
- administrative records with monitor and interviewer characteristics.

These data allowed for operationalization of multiple sources of variation in monitors' error detection that are not captured in the monitoring system and results.

Section 2 describes the data compiled and analysis used to assess what factors explain variation in monitors' error detection. ***Section 3*** provides the results from the analysis. ***Section 4*** summarizes key conclusions from this investigation, discusses the implications of the results for telephone survey monitoring in practice, and suggests further research to improve understanding of how various factors contribute to variation in interviewer error detection.

2. Description of Data and Analysis

2.1 Data

This section describes compiled and analysis used to assess what factors explain variation in monitors' error detection.

2.1.1 QUEST monitoring sessions

The QUEST data for this investigation came from an ongoing national ABS telephone survey on community health issues. The data collection period for this study was October

2013 through August 2014. Two key assumptions applied to these QUEST monitoring data were (1) the same survey instrument and protocol was maintained throughout the field period and (2) monitoring sessions were assigned randomly to monitoring team members. These assumptions appear to have been maintained for the data collection period examined. The QUEST session data from October 4, 2013 and April 14, 2014 included the following variables for 907 monitoring sessions conducted by 11 monitors and including 49 interviewers:

- session ID number,
- monitor ID number,
- interviewer ID number,
- indicator for live versus recorded monitoring session, and
- number of interviewing errors detected in the session for each interviewing skill area.

Each of these 907 sessions involved either a complete or partial interview, based on the study outcome codes. Because some of the skill areas – authenticity, case management skills, and professional behavior – had very few errors detected across the sessions, the 10 skill areas were combined into two continuous composite outcomes measures as follows:

> ***Interview administration skills*** = case management + keying + question administration + probing + feedback + protocol
>
> ***Professional and interpersonal behaviors*** = authenticity + initial contact + presentation skills + professional behavior

These two composites based on monitors' assessments served as the outcome variables in the models estimated.

2.1.2 Responses to a survey of monitors

This survey collected data from all call center monitors and the survey data from 11 monitors who conducted any of the 907 monitoring sessions were incorporated into the analysis. The response rate to the monitor survey was 100%, using AAPOR RR1. The survey asked monitors to answer 10 questions about their orientation to the following five aspects of their work:

1. accurately identifying all interviewer errors in a session,
2. correctly entering all interviewer errors detected,
3. navigating through the QUEST application,
4. receiving the training needed to successfully conduct sessions, and
5. receiving the supervision needed to be successfully conduct sessions.

For each of these five dimensions, the survey asked monitors to answer separate questions based on whether they were conducting a *live* monitoring session versus a *recorded* monitoring session. Because monitor responses were highly correlated across

items a composite measure was created by standardizing the response values and combining all responses into a single variable for each monitor.[1]

2.1.3 Administrative data with monitor and interviewer characteristics

Data from RTI personnel records included the following characteristics of monitors and interviewers who worked on the study:

- experience, measured by months of employment in RTI's call center,
- telephone interviewing experience prior to working in RTI's call center,
- highest education level completed, and
- gender.

For the second outcome, monitors' detection of protocol violations for professional and interpersonal behaviors, model 3 excluded monitor gender, as including this covariate prevented the model from converging.

2.2 Analysis

For the two outcome variables, the continuous composite monitor ratings for interview administration skills and professional and interpersonal behaviors, the following three generalized linear models were produced:

***Model 1*:** Estimation of a baseline model with monitor and interviewer as random effects, without any covariates. This model allowed for calculation of the variance attributable to monitors in both outcome variables, net of interviewer effects.

***Model 2*:** Estimation of monitor and interviewer as random effects with the monitor survey composite added as a categorical fixed covariate.

***Model 3*:** Estimation of monitor and interviewer as random effects with both the monitor survey composite added as a categorical fixed covariate and live vs. recorded session, monitor characteristics, and interviewer characteristics added as fixed covariates.

For the second outcome, and professional and interpersonal behaviors, model 3 excluded monitor gender, as including this covariate prevented the model from converging.

3. Results

3.1 Model 1

As shown in ***Tables 1*** and ***2***, model 1 produced estimates only for monitor and interviewer as random effects (without any covariates) for both interview administration skills and professional and interpersonal behaviors. This model allowed for calculation of the variance attributable to monitors in both outcome variables, net of interviewer variance. Dividing the monitor variance by the monitor variance plus the residual variance for each of the two outcomes produced a monitor variance ratio for each outcome. The monitor variance ratio was greater for interview administration skills (0.030) than for professional and interpersonal behaviors (0.005).

[1] A copy of the monitor survey is available from the lead author.

The monitor variance ratio mirrors the formula for Kish's *rho* $_{\text{interviewer}}$, a unit-free statistic for the intra-interviewer correlation associated with interviewers. As such, the monitor variance ratios for interview administration skills and professional and interpersonal behaviors represent the intra-monitor correlation associated with monitors for each set of interviewing errors observed. The impact of monitor variance on error detection rates can therefore be treated as a design effect, where the average session workload is an important factor. The average session workload for monitors in these QUEST data was 907 sessions/11 monitors = 82.45 sessions. The *deff* $_{\text{monitor}}$ for interview administration skills was 3.430 and the *deff* $_{\text{monitor}}$ for professional and interpersonal behaviors was 1.405. These design effect calculations indicate the combination of relatively high workloads and a relatively high monitor variance ratio appeared to have a significant impact on detection of interview administration skill errors, but the significantly lower monitor variance ratio for professional and interpersonal behaviors resulted in a smaller effect.[2]

Table 1: Model 1 Parameter Estimates for Interview Administration Skills

Parameter	Subject	Estimate	Std. Error
Intercept	Monitor	0.0108	0.0070
Intercept	Interviewer	0.0530	0.0160
Residual		0.3732	0.0181

Table 2: Model 1 Parameter Estimates for Professional and Interpersonal Behaviors

Parameter	Subject	Estimate	Std. Error
Intercept	Monitor	0.0002	0.0003
Intercept	Interviewer	0.0003	0.0005
Residual		0.0431	0.0021

3.2 Model 2

Model 2 produced estimates only for monitor and interviewer as random effects with the monitor survey composite added as a categorical fixed covariate. Results for this model are shown in ***Tables 3a-c*** and ***4a-c***. For both interview administration skills and professional and interpersonal behaviors, the monitor survey composite did not have any significant association with error detection outcomes.

[2] The observed interviewer variances could have been inflated due to cases not being perfectly randomized (interpenetrated) among the 49 interviewers. Given the focus was on monitor variability, no attempt was made to remove the "assignment effects" from the interviewer variance estimates.

Table 3a: Model 2 Parameter Estimates for Interview Administration Skills

Parameter	Subject	Estimate	Std. Error
Intercept	Monitor	0.0126	0.0089
Intercept	Interviewer	0.0532	0.0161
Residual		0.3734	0.0181

Table 3b: Model 2 Solutions for Fixed Effects for Interview Administration Skills

Effect	Estimate	Std. Error	DF	t value	Prob > |t|
Intercept	0.2381	0.0766	8	3.11	0.0145
Survey Composite Low	0.0791	0.1338	848	0.59	0.5542
Survey Composite Med.	-0.0025	0.0929	848	-0.03	0.9786

Table 3c: Model 2 Type III Tests of Fixed Effects for Interview Administration Skills

Effect	Num DF	Den DF	t value	Prob > |t|
Survey Composite	2	848	0.21	0.809

Table 4a: Model 2 Parameter Estimates for Professional and Interpersonal Behaviors

Parameter	Subject	Estimate	Std. Error
Intercept	Monitor	0.0000	0.0003
Intercept	Interviewer	0.0002	0.0005
Residual		0.0432	0.0021

Table 4b: Model 2 Solutions for Fixed Effects for Professional and Interpersonal Behaviors

Effect	Estimate	Std. Error	DF	t value	Prob > |t|
Intercept	0.0530	0.0117	8	4.50	0.0020
Survey Composite Low	0.0143	0.0274	848	0.52	0.6016
Survey Composite Med.	-0.0245	0.0153	848	-1.63	0.1044

Table 4c: Model 2 Type III Tests of Fixed Effects for Professional and Interpersonal Behaviors

Effect	Num DF	Den DF	t value	Prob > |t|
Survey Composite	2	848	1.92	0.147

3.3 Model 3

Model 3 added additional fixed covariates to model 2, including live vs. recorded session, monitor characteristics, and interviewer characteristics. The results for interview administration skills presented in ***Table 5*** indicate:

- Live sessions were associated with a lower probability of interview administration errors being observed than in recorded sessions.
- Greater monitor experience was associated with a lower probability of observing interview administration errors than less monitor experience.
- Monitor gender was not quite significant, but this result suggests male monitors could be associated with a higher probability of observing interview administration errors than female monitors.

The results for professional and interpersonal behaviors presented in ***Table 6*** indicate:

- As for interview administration errors, live sessions were associated with a lower probability of professional and interpersonal behavior errors being observed than in recorded sessions.
- Male interviewers were associated with a higher probability of observing professional and interpersonal behavior errors than female interviewers.

Interviewer experience was not quite significant, but result suggests greater interviewer experience could be associated with a lower probability of observing professional and interpersonal behavior errors.

Table 5: Model 3 Type III Tests of Fixed Effects for All Covariates for Interview Administration Skills

Effect	Num DF	Den DF	F value	Prob > |F|
Live vs. Recorded	1	846	33.07	< 0.001
Interviewer Gender	1	846	0.02	0.896
Interviewer Experience	1	846	2.16	0.142
Interviewer Education	3	846	2.23	0.084
Monitor Gender	1	846	3.45	0.064
Monitor Experience	1	846	4.12	0.043
Monitor Education	3	846	0.35	0.789
Survey Composite	2	846	1.44	0.238

Table 6: Model 3 Type III Tests of Fixed Effects for All Covariates for Professional and Interpersonal Behaviors

Effect	Num DF	Den DF	F value	Prob > |F|
Live vs. Recorded	1	846	8.32	0.004
Interviewer Gender	1	846	5.06	0.025
Interviewer Experience	1	846	3.74	0.053
Interviewer Education	3	846	1.2	0.310
Monitor Experience	1	846	0.22	0.642
Monitor Education	3	846	0.55	0.647
Survey Composite	2	846	0.34	0.714

4. Conclusions and Implications

This investigation of multiple factors associated with monitor variability in detecting interviewing errors provided three main conclusions. First, variance attributable to monitors appeared to be non-trivial for detection of errors for interview administration skills, but relatively small for error detection for professional and interpersonal behaviors. The generally high work load among monitors in the data examined (about 82 sessions on average) combined with a relatively high monitor variance ratio appeared to have a significant impact on detection of interview administration skill errors in this study. The substantially lower monitor variance ratio for professional and interpersonal behaviors appeared to result in little impact of monitor variance on error detection for this outcome. The smaller variance associated with monitors' detection of protocol violations for professional and interpersonal behaviors appeared to be driven mostly by the consistently high ratings monitors assigned to these skill areas. Furthermore, interview administration

skills included more high-level interviewing skills which are more difficult for monitors to assess, even in a highly standardized monitoring system. This distinction was likely a factor in explaining why monitors' detection of protocol violations for these higher-level interviewing skills varied to a greater degree.

Second, monitors' degree of confidence in skills, training, and supervision was not significantly associated with error detection. This could reflect limitations of the survey measures, a lack of true impact of monitors' orientations on the monitoring results, or some combination of both factors.

Third, most other of the other factors examined did not account for significant variation in detection of either interview administration skills errors or professional and interpersonal behavior errors, with three exceptions:

1. For both outcomes, live monitoring sessions were associated with a lower probability of observed errors. This consistent finding suggests one clear value of recorded monitoring sessions is allowing sufficient time for monitors to accurately capture interviewing errors. Although live monitoring sessions have value for observing telephone interviewing errors as they occur, this finding suggests the interview pace or other complications could limit monitors' ability to accurately capture all interviewing errors.

2. For interview administration skills, greater monitor experience was associated with a lower probability of observing errors. This finding is somewhat counter-intuitive, and deserves further attention. More experienced monitors generally would be expected to have greater ability to detect and record protocol violations errors. Unmeasured factors such as relationships with current interviewers or previous experience working as an interviewer could have affected the judgment of more experienced monitors to a greater degree than those with less experience.

3. For professional and interpersonal behaviors, male interviewers were associated with a higher probability of observing errors. All male monitors included in this analysis are bilingual monitors (English and Spanish). The potential impact of differences by language could provide at least a partial explanation of this finding and should be examined more closely in future research.

As suggested in prior sections of this paper, this research had the following notable limitations:

- The data set was ultimately too small to detect the extent and sources of monitor variability specified by the study goals. Although over 900 monitoring sessions were included in the analysis, as noted in ***Section 2.1.1***, these sessions included only 11 monitors and 49 interviewers. A significantly larger data set could have allowed for examination of all, or nearly all, of the individual skill areas as outcome variables. For future research, a power analysis is needed to determine the number of sessions, monitors, and interviewers needed to meet the analytic goals.

- Likewise, the need to use composite measures for the two key outcome variables could have obscured more specialized effects. The degree of monitor variation in detecting protocol violations could have differed significantly across the specific skill areas combined into each composite measure. Being able to examine most

or all of the specific skill areas as individual outcomes would have provided a more precise assessment of where, how much, and, perhaps, why monitors varied in detecting interviewer errors.

- A richer set of monitoring explanatory variables would have likely improved the interpretation of the findings. As noted in ***Section 2.1.2***, monitor responses to the survey questions were highly correlated across items, so a single composite measure was created. This measure was not significantly associated with either of the two outcomes, and therefore contributed little to the analysis. Improved measurement of monitors' orientations relevant to their work could explain some variation in the observed results, such as differences between more and less experienced monitors.

These limitations clearly indicate further research is needed to improve our understanding of how various factors contribute to monitors' variation in detecting interviewer errors, such as:

- Estimating the same models with the full set of monitoring sessions from the completed study or another study with a larger data set could increase the number and range of observed errors in the data set and allow for a more thorough analysis of monitors' error detection rates for specific skill areas.

- Conducting the same analysis of monitoring session data for additional studies, to assess the generalizability of these findings. These findings would suggest other telephone surveys using the same monitoring protocol, and same types of monitors, would show similar monitor variations. Data could also be combined across a number of surveys to determine how survey-dependent the results are.

- Improving the monitor survey to better capture monitors' orientations relevant to their work might be useful. The limited survey literature on survey quality monitoring poses a challenge to capturing monitor orientations likely to be related to how and when they observe interviewing errors. Having a clearer sense of whether and how monitor orientations to their job might play a role in their detection of protocol violations could suggest ways to improve monitor training or further standardize the monitoring system.

Acknowledgements

The authors thank other current members of RTI's QUEST team for their contributions, including Susan Kinsey (Lead), Melissa Cominole, Dave Foster, Sridevi Sattaluri, Curry Spain, and Howard Speizer.

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. P., Herget, D., Morton, J., & Willis, G. (2001). The feasibility of monitoring field interviewer performance using computer audio recorded interviewing (CARI). *Proceedings of the American Statistical Association, Section on Survey Research Methods,* pp. 1068–1073.

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.

Currivan, D. B., Stone, D. B., Spain, C. J., & Tate, N. M. (2013. *Variability in error detection among telephone monitors*. Presented at American Association for Public Opinion Research annual conference, Boston, MA.

Durrant, G.B., Groves, R.M., Staetsky, L. and Steele, F. (2010). "Effects of Interviewer Attitudes and Behaviors on Refusal in Household Surveys." *Public Opinion Quarterly* 74: 1–36

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

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. (2014). Computer audio-recorded interviewing as a tool for survey research. *Social Science Computer Review*, *32*(1) 90–104. http://ssc.sagepub.com/content/early/2013/07/26/0894439313500128.abstract

Thissen, M. R., Park, H., & Nguyen, M. (2013). Computer audio recording: A practical technology for improving survey quality. *Survey Practice*, *6*(2), 1–7. http://www.surveypractice.org/index.php/SurveyPractice/article/view/38

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 .