Assessing the Reliability of Conversational Interviewing November2016

William Mockovak
Bureau of Labor Statistics, 2 Massachusetts Ave, N.E., Washington, DC 20212

Abstract
Although not widely used, advocates of conversational interviewing argue that it reduces respondent burden and results in higher quality data in situations where the survey information being requested is complex, highly sensitive, or where the topic elicits stress. However, because it does not rely on a scripted interview, conversational interviewing also raises concerns and challenges concerning its consistency and reliability. This study explored an approach for assessing the reliability of response coding in a conversational interview by asking 86 interviewers to observe and code a video of an interview conducted conversationally. Responses coded by interviewers were then compared to gold-standard answers. To assess the impact of experience, the video observations were conducted on two occasions: once immediately after initial training, and, again, four months later in the data collection period. Results after initial training showed high levels of interviewer consistency using percent agreement with the gold-standard answers, and after four months of data-collection experience, performance significantly improved. Other measures of reliability that correct for chance agreement such as Intraclass Correlation Coefficients (ICC) and Cronbach's alpha were also explored and their use as possible tools for assessing interviewer consistency discussed.

Key Words: Conversational interviewing, interviewer reliability, interviewer consistency, calibration training, interviewer performance

1. Introduction

Although conversational interviewing is not widely used in survey research, some survey managers have decided that there are situations where it is more likely to produce higher quality data than standardized interviewing. For example, in situations where respondents believe that the information being requested is highly sensitive, where complex information is being collected (for example, health insurance plans), or where the survey topic elicits stress (for example, discussing sexual practices, reasons for job layoffs). Therefore, interviewers in these types of situations have used flexible, conversational interviewing methods to relax respondents, encourage open conversations, improve response, address respondent concerns, and ideally obtain high quality data.

Definitions of conversational interviewing vary and the approach can be implemented to varying degrees; therefore, implementation can be viewed as existing on a continuum. As one example, Schober and Conrad (2004) describe a limited, experimental implementation of conversational interviewing that allowed interviewers to use an improvised explanation to provide clarifications for selected survey questions after the questions had been asked as worded. Specifically, interviewers were allowed to use an unscripted definition based on a factual definition when they judged it necessary, for example, when a respondent appeared confused. This approach improved data quality for survey questions with more complicated conceptual mappings, but at the cost of a longer interview. On the other hand, Mockovak (2010) describes a situation where all data items had been collected

conversationally for years, but where more standardization was introduced in the form of a structured checklist.

The advantages and disadvantages of standardized and conversational interviewing have been discussed elsewhere (Suchman and Jordan, 1990; Mishler, 1986; Beatty, 1995; Maynard and Schaeffer, 2002). This paper looks at the use of conversational interviewing in an establishment survey, the Occupational Requirements Survey (ORS), which the Bureau of Labor Statistics (BLS) is conducting for the Social Security Administration (SSA). Ferguson (2013) describes the objectives of ORS, but the survey collects detailed information on a wide variety of occupational characteristics including the following topics:

- physical demand characteristics/factors of occupations (e.g., strength, hearing, or stooping);
- specific vocational preparation requirements, which include educational requirements, prior experience, licensing and certification, and post-employment training;
- mental and cognitive demands of work;
- environmental conditions in which the work is completed; and
- descriptions of primary tasks performed in an occupation.

Survey managers decided to encourage the use of conversational interviewing based on early field testing for ORS in which the data collectors (Field Economists or FEs) were required to use standardized interviewing procedures without any deviations, following guidelines described in Fowler and Mangione (1990). These early field tests produced discouraging results. Respondent, interviewer, and observer debriefings consistently and clearly indicated that both respondents and FEs did not like the fully standardized approach finding it boring, repetitious, lengthy, and tedious, with possible negative effects on data quality and respondent cooperation. As Handel (2015) notes when reviewing ORS pretest findings, "standardized interviews and surveys limit interviewer variance, including possible halo effects, but they also limit the interviewer's ability to probe for details, provide helpful prompts tailored to specific situations, and identify and correct respondents' misunderstandings and biases. Indeed, respondents themselves reportedly felt conversational interviews improved the quality of data and eased the interview process."

Since those early tests the ORS interview has only gotten even more demanding with about 71 data elements, plus lists of primary job tasks, being collected for anywhere from two to eight occupations during an interview in the same establishment. Based on data from a pre-production (dress rehearsal) test, an average ORS interview lasts about 60 minutes.

1.1. Characteristics of Conversational Interviewing as Implemented in ORS

Since conversational interviewing can vary in how it is implemented, some of the guidelines that have been used to implement conversational interviewing in ORS are listed below:

- The data collectors, known as Field Economists (FEs), are trained to thoroughly research business establishments before contacting them to learn what their primary mission is, what types of occupations they employ, and to search for materials related to the upcoming data collection (for example, written position

descriptions). FEs are expected to be knowledgeable about the primary business activities and employee base before the start of the interview.

- Although many initial contacts with establishments will be conducted by phone (for example, to introduce the survey and to set up appointments), a significant part of the data collection occurs face-to-face. However, after an initial contact the FEs can communicate and collect additional or missing data using email or the phone. For example, respondents might send interview-relevant information such as written position descriptions to FEs using email.

- During an interview, FEs are encouraged to ask respondents if they did not understand a question and to offer clarifications and elaborations as necessary.

- Different sections of the interview can be addressed in different orders, usually depending on the preference of a respondent or how topics segue. Moreover, other time saving approaches can be used. For example, if two office workers are sampled in an establishment, the physical and environmental demand questions can be asked at the same time for these two workers, whereas if an office and blue-collar job were sampled, these questions might be asked separately.

- Task lists, or descriptions of key tasks in a job, are collected using an open-ended question.

- When feasible, job observations can be set up to collect information and to verify information provided by a respondent.

- FEs receive extensive on-the-job training and work is reviewed through a comprehensive quality assurance process.

1.2. Interviewer (Field Economist) Selection and Training

The traditional implementation of standardized interviewing requires interviewers to read the questions exactly as worded with no deviations, to use only nondirective probing, and to maintain a neutral interviewing approach with no discretion involved in how answers are recorded. In addition, Fowler and Mangione (1990) call for practice-oriented training, systematic monitoring of interviews, and close supervision. Since the basic task is to read the questions and probing is very limited, one advantage of the standardized approach is that it leads to simplified training, which means that money can be saved through shorter training and the selection of less skilled interviewers.

In contrast, conversational interviewing requires interviewers to be constantly alert to possible issues with respect to respondent comprehension so that explanations can be tailored to individual respondents or questions even rephrased, if necessary. To accomplish such interventions successfully, interviewers are expected to have expert knowledge of the intent of each question so that an immediate determination can be made if the information being provided by respondents is accurate and meets objectives.

To accomplish this high level of interviewing, more stringent selection criteria are applied. Interviewers (Field Economists) must have a college diploma with a required number of economic or statistics courses, and are required to complete a rigorous training and certification program before being allowed to collect data independently (Ferguson, 2013). Initial training is also supplemented by expert mentoring and monitoring to ensure that data collectors can work independently. Finally, an extensive quality control program exists to ensure that data quality meets standards (Harney and Brown, 2015; Meharenna, 2015).

1.2. Quality Control

Once a survey is in the field, several approaches can be used to monitor the quality of interviewing. Some possible approaches are reinterview, field observations or monitoring (especially in phone centers), quality reviews of completed questionnaires, and analysis of paradata or exploratory analyses of actual data. This paper discusses use of another option, called calibration training (Mockovak et al., 2015), which is a type of refresher training but with the following characteristics:

- Training is focused on a specific skill identified through quality reviews, observations, or analysis of paradata or actual data.
- Sessions are limited in duration (generally three hours), and are often conducted online (remotely) to accommodate interviewers working in dispersed geographic areas.
- Some type of pre- vs post-assessment activity is used to measure the effects of the intervention (often training about a specific topic).
- Performance is measured quantitatively against previously determined gold standards.
- Feedback on performance is provided to the interviewers.

A key benefit of calibration training is that results from these sessions can quickly lead to changes in procedures, survey definitions, initial training, or quality control activities.

1.3 Research Objectives

The key research objectives in this study were the following:

- How reliably and accurately can Field Economists code survey responses in a conversational interview that they observe?
- What impact does four months of experience conducting conversational interviews have on their performance?

2. Methodology

One possible approach for assessing the quality of conversational interviewing would be to ask each interviewer to conduct a conversational interview and then assess how accurately the data were collected. Unfortunately, such an approach poses a wide variety of logistical challenges (for example, obtaining enough "respondents," ensuring equal difficulty of the interviews, determining gold-standard answers, etc.) and is extremely time and labor intensive, especially when the interviewers are widely dispersed geographically. Therefore, the decision was made to create two videos (pre and post) that showed a complete interview being completed using a conversational approach. FEs (interviewers) participating in the calibration training session were then asked to view the videos and code the responses. Figure 1 shows the approach used in this study.

Figure 1: Approach Used for the Calibration Evaluation
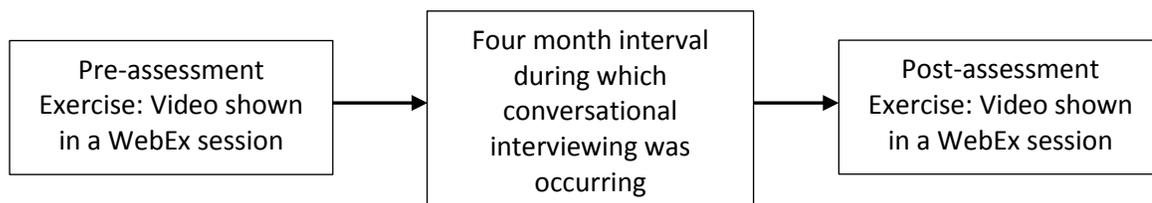
107 Field Economists participated in the pre-assessment exercise, which was scheduled in the first month after initial training had been conducted, and 92 participated in the post-assessment exercise, which was held four months later. To accommodate FE schedules, multiple sessions were held during two different time spans:

- "Pre" Session   January 19-30, 2015
- "Post" Session   May 14-19, 2015

The purpose of the pre-session was to establish a baseline for future comparisons of inter-rater agreement and reliability. To ensure that the pre- and post-course videos would provide a valid, comparable measure of an FE's coding skills, the following steps were taken:

1. The establishment information (name, address) in the "pre" and "post" sessions differed and the selected occupations had different job titles, but all key data elements such as educational, cognitive, physical, and environmental demands were identical in both videos.

2. Each of the videos showed a conversational interview of one occupation (computer support position); however, the Field Economists (interviewers) and respondents in the pre and post videos were represented by different actors (the FEs and respondents in the videos were highly trained volunteers).

3. The pre- and post-assessment calibration sessions were spaced approximately four months apart to determine how experience conducting conversational interviewing affected FEs' performance (in the intervening four months, FEs were collecting ORS data in a field test). A four-month interval was chosen because it was judged to be a long enough time to have an impact on performance, and also a long enough time for participants to forget the occupational details that had been covered in the pre-assessment video.

While observing the videos, participants could use whatever tool or tools they preferred to use to record data (printed data collection forms were available), but they were then required to enter their answers in an Adobe fillable form that had been emailed to them prior to the session. FEs were required to submit their answers online within one hour after the completion of the WebEx session. There was no discussion during the WebEx session, and FEs were also told to save a copy of the form so that they could review the gold-standard answers when they were sent out after the training sessions had been conducted. Other than general introductions, the only activity that occurred during the online WebEx sessions was showing the relevant video and dealing with occasional technical problems that FEs experienced.

Because a conversational interview does not follow any specific topic order, FEs must be prepared to address any ORS element at any time, which can be a very demanding task. In the approach used in this study, the observing FE did not control the flow of the interview, which differs from how a real-life interview would be conducted. Also, occasional technical difficulties can occur within WebEx that make the task even more difficult (for example, poor quality audio, videos that freeze). To address this assortment of anticipated difficulties, each video had pauses inserted at pre-determined spots so that there was time for FEs to take notes and code their answers. The length of these pauses was fixed in each calibration session. Each video was watched once with these pre-determined pauses to allow note taking, and then a second time with no pauses. Each WebEx session lasted about an hour.

As mentioned, FEs were told to submit their answers within one hour after the conclusion of the training session. If desired, FEs could use this time to review their coding and to check technical guidance (e.g., Collection Manual, Tech Memo, training materials) if they had a procedural question. This approach mirrors what FEs might do after a typical interview.

2.1 Technical Difficulties Encountered during the WebEx Sessions

Some technical difficulties were encountered in WebEx during some sessions that could negatively impact FE performance. Fortunately, these difficulties occurred for a relatively small number of FEs. For example, some FEs reported difficulty hearing the audio that accompanied the video, and on rare occasions, the video picture itself disappeared, although the audio continued. When these types of difficulties occurred, work-arounds were immediately implemented.

2.2 Analysis Plan

After all the sessions had been completed, FEs' answers were compared to gold-standard answers that had been previously developed by an internal team of experts (the experts used a consensus approach to agree on responses for individual items). Some of the ORS elements are dichotomous; that is, they indicate either the presence or absence of an element such as the requirement to drive or use a keyboard, whereas, other elements require the collection of duration using the following scale:

- Seldom: up to 2 percent of the workday
- Occasional: from 2 percent to one-third of a workday
- Frequent: from one-third to two-thirds of a workday
- Constant: two-thirds or more of a workday

In some instances, respondents may assert that a requirement is present but the duration of the activity is unknown; these requirements are captured as "present, duration unknown." Each entry for each data element, whether dichotomous or a value on the duration scale, was coded as either correct or incorrect.

3. Results

3.1. Overall Performance: Percent Agreement

Percent agreement is not an ideal measure of FE consistency because it does not account for the role of chance in obtaining the correct answer, but it provides an intuitive measure of performance that is easily understood, and it can be used to quickly identify elements that seem to be causing difficulties. Also, as previously mentioned, answers are being compared against gold-standard answers. Of course, due to the technical difficulties that some participants encountered during the calibration session, 100 percent agreement on any element should be rare.

Table 1 shows the percent agreement among the 86 participants who could be matched between the pre- and post-assessment exercises (the number of attendees between the pre and post sessions varied because some attendees were office staff who were only indirectly involved in data collection).

Results are shown for individual elements, sections of the ORS form, and an average is computed for the entire form. Overall percent agreement increased from 88.9 percent for the "pre" scores to 92.1 percent for the "post" scores, or a 3.2 percent increase. This difference was significant using a paired-samples t-test (t= -3.093, df= 85, p= .003).

Table 1. Mean Proportion of Agreement on ORS Data Elements on the Pre- and
Post-Assessment Activities

ORS Data Element	Pre	Post	Difference Post-Pre
Job Details			
Job Title	1.000	1.000	0
Occupational Employment	.872	.942	.07
Work Schedule	.970	.814	-0.156
SOC	.721	.593	-0.128
FT/PT	.977	.977	0
Time/Incentive	.970	.977	.007
Position	.837	.884	.047
Mean	0.901	0.884	-0.023
Educational Requirements			
Degree	.988	.814	-0.174
Prior Work Experience	.977	.970	-0.007
Post Work Experience	.907	.930	.023
Certification, Licenses, Training	.970	.860	-0.11
Mean	0.961	0.894	-0.067
Cognitive Elements			
Complicated	.970	.977	.007
Controlled	.970	.977	.007
Deviations in Work Tasks	.953	.860	-.093
Deviations in Work Schedule	.970	.970	0
Deviations in Work Location	.988	.988	0
Regular Contacts Verbal Interaction	.953	1.000	.047
Regular Contacts Type of Interaction	.826	.744	-0.082
Other Contacts Verbal Interaction	.942	.872	-0.07
Other Contacts Type of Interaction	.744	.651	-0.093
Mean	0.924	0.893	-0.031
Physical Demands			
Sit (duration)	.907	.953	.046
Stand/walk (duration)	.942	.970	.028
Sit/stand at will (presence/absence)	.884	.988	.104
Most weight	.970	.977	.007
Weight: More than 2/3 of time	.372	.558	.186
Weight: 1/3 to 2/3 of time	.872	.942	.07
Weight: 2% to 1/3 of time	.953	.970	.017
Weight: <2% (seldom)	.884	.849	-0.035

ORS Data Element	Pre	Post	Difference Post-Pre
Push/Pull: hand/arm	.953	.942	-0.011
Push/Pull: foot/leg	.977	.970	-0.007
Push/Pull: foot only	.977	.953	-0.024
Reaching: Overhead (duration)	.970	.953	-0.017
Reaching: At/below shoulder	.895	.930	.035
Reaching: At/below shoulder (one or both)	.791	.814	.023
Gross manipulation (duration)	.314	.895	.581
Gross manipulation (one/both)	.174	.872	.698
Fine manipulation (duration)	.884	.919	.035
Fine manipulation (one/both)	.826	.628	-0.198
Foot/leg controls	.977	.970	-0.007
Keyboarding: Traditional	.970	.977	.007
Keyboarding: 10-key	.942	1.000	.058
Keyboarding: Touch screen	.970	1.000	.03
Keyboarding: Other	.953	.988	.035
Stooping (duration)	.953	.953	0
Crouching (duration)	.674	.663	-0.011
Kneeling (duration)	.826	.814	-0.012
Crawling (duration)	.919	.919	0
Climbing: Ramps/stairs, job-related (duration)	.837	.953	.116
Climbing: Ramps/stairs, structure (duration)	.895	.942	.047
Ladders, ropes, scaffolds	.802	.970	.168
Driving (yes/no)	.988	.988	0
Mean	0.847	0.910	0.064
Communication Requirements			
Communicating verbally	.826	.895	.069
Hearing test	.977	.988	.011
Hearing: 1 on 1	.977	.988	.011
Hearing: in group	.977	.988	.011
Hearing: on phone	.970	.988	.018
Hearing: other sounds	.919	.814	-0.105
Near visual acuity	.988	.988	0
Far visual acuity	.970	.970	0
Peripheral vision	.907	.953	.046
Mean	0.946	0.952	0.007
Environmental Conditions			
Noise intensity level	.663	.953	.29
Outdoors (duration)	.953	.988	.035
Extreme heat (duration)	.970	.988	.018
Extreme cold (duration)	.970	.988	.018

ORS Data Element	Pre	Post	Difference Post-Pre
Wetness (duration)	.930	.988	.058
Humidity (duration)	.953	.988	.035
Heavy vibration (duration)	.942	.988	.046
Fumes, noxious odors, dusts, gases (duration)	.907	.988	.081
Toxic, caustic chemicals (duration)	.919	.988	.069
Proximity to mechanical moving parts (duration)	.907	.988	.081
High exposed places (duration)	.884	.953	.069
Mean	0.909	0.982	0.073
Average across all sections	0.889	0.921	0.032

As shown in Table 2, percent agreement varied by section, with the largest average differences between pre and post (both positive and negative) occurring in the Educational Requirements, Physical Demands, and Environmental Conditions sections.

Table 2. Percent Agreement by Section (N=86)

Section	Pre	Post	Post-Pre
Job Details	90.1%	88.4%	-2.3%
Educational Requirements	96.1%	89.4%	-6.7%
Cognitive Elements	92.4%	89.3%	-3.1%
Physical Demands	84.7%	91.0%	6.4%
Communication Requirements	94.6%	95.2%	0.7%
Environmental Conditions	90.9%	98.2%	7.3%
Overall	88.9%	92.1%	3.2%

Overall, the number of actual elements correctly identified significantly increased from the pre-test to the post-test (see Table 3). The Physical Demands' and Environment Conditions' sub-sections saw significant improvements in the number of elements correctly identified; however, there was a significant decrease in the number of elements correctly identified in the Educational Requirements section.

Table 3. Average Number of Elements Correctly Identified by Section

Section	Pre	Post	Post-Pre	t (paired)	df	p value
Job Details	6.34	6.19	-0.15	-1.17	85	0.24
Educational Requirements	3.84	3.57	-0.27	-3.21	85	<.01*
Cognitive Elements	8.02	8.26	0.24	0.81	85	0.42
Physical Demands	26.23	28.2	1.97	5.86	85	<.01*
Communication Requirements	8.50	8.57	0.07	0.41	85	.68

Environmental Conditions	9.99	10.80	0.81	3.10	85	<.01*
Overall	63.20	65.36	2.16	2.92	85	< .01*

3.2 Other Measures of Inter-Rater Reliability

As mentioned, one problem with using percent agreement as a measure of accuracy is that it does not account for the role of chance. Measures of inter-rater reliability such as Cohen's kappa can be used to adjust for chance agreement of categorical judgments, and these measures have been generalized for use with multiple raters. Although kappa can be negative, its values generally fall between zero and one, with a zero value indicating that judges did not agree with each other any more than would be predicted by chance. However, kappa can have serious problems under certain circumstances, most notably, a phenomenon called the "paradox of kappa" can occur. This paradox occurs when absolute agreement is very high on nominally scaled data, but, contrary to expectations, very low levels of estimated reliability result (Gwet, 2008; Warrens, 2010). As shown in Table 1, there are very high levels of agreement on nominally scaled data, so it was determined that computing kappa values would provide misleading measures of inter-rater reliability.

3.3. Internal Scale Reliability

Another way to look at reliability is in terms of the inter-correlations of the items, which provide a measure of internal consistency, or how closely related a set of items are as a group. Internal reliability is important because it's a prerequisite for a valid measure.

A measure that is commonly used to assess scale reliability is Cronbach's alpha coefficient. This measure can be thought of as the expected correlation of two tests that measure the same construct. A useful property of Cronbach's alpha is that it produces a single consistency estimate of inter-rater reliability across multiple judges. The next table shows the value of Cronbach's alpha for the full ORS instrument and most major sub-sections for the pre- and post-assessment activities.

Table 4. Cronbach's Alpha Based on Standardized Items

ORS Section	Pre	Post
Job Details	.497	.411
Education/training	.450	.290
Cognitive demands	.647	.458
Communicative demands	.850	.658
Physical demands	.894	.660
Environmental conditions	.964	.986
Full Interview	.931	.772

A reliability coefficient of 0.70 or higher is considered "acceptable" in most social science research situations.) [1] Although the "Full Interview" values can be considered acceptable to excellent, these results generally show lower internal reliability for the post-course measure than for the pre-course measure, and low internal reliability for some sections, especially in the post-assessment.

---

[1] http://www.ats.ucla.edu/stat/spss/faq/alpha.html

However, these results can also be misleading. For example, although Coefficient Alpha for the physical demands section in Table 4 dropped between the pre and post measurements, the mean number of items answered correctly actually increased from 84.7 percent to 91 percent (see Table 1).

3.4. Results for the List of Critical Job Tasks

An important part of the descriptive information obtained in an interview about each job is a description of the task list, or the job's most important tasks or activities. A task list is an important component of data collection because it provides an overview of the requirements of a job, which an FE can then use as context when collecting more detailed job requirements data. These lists are collected conversationally through an open-ended question.

As mentioned in the Methodology section, the occupations used in the pre-course and post-course assessments were computer-support jobs with different titles, but identical task lists. The gold-standard answers were determined in advance by a team of experts (using a consensus approach), and consisted of the following eight major tasks (order is not important):

1. Provides computer support for employees.
2. Installs hardware or software.
3. Ensures compatibility/integration.
4. Tests network.
5. Troubleshoots software & network issues.
6. Serves as a backup to Network Analysts.
7. Examines IT needs, by asking questions, participating in focus groups.
8. Runs network cables.

Although FEs were encouraged to write task lists similar to the targeted list shown above (i.e., start each task with an active verb, etc.), the actual lists generated by FEs differed in structure and clarity. Therefore, judgment was required to determine if a task had been adequately described or not, so two coders independently coded the task lists created by FEs in the pre-test and post-test. Coder 1 remained the same in the pre-assessment and post-assessment, but Coder 2 changed. The next table shows how often coders correctly identified each of the eight tasks as being present and the level of agreement between the two coders.

Table 5. Average Number of Times a Task Was Identified as Being Present by a Coder & Percent Agreement between Raters

Coder	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8
Pre-Assessment (N=97)								
Coder 1	0.97	0.78	0.68	0.81	0.90	0.62	0.87	0.07
Coder 2	0.95	0.72	0.61	0.74	0.76	0.60	0.87	0.06

Coder	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8
% agreement between Coder 1 and Coder 2	95.9%	85.6%	84.5%	91.2%	82.5%	78.4%	97.9%	99.0%
Post- Assessment (N=76)								
Coder 1	0.92	0.99	0.82	0.59	0.91	0.78	0.80	0.00
Coder 2	0.80	0.99	0.84	0.51	0.90	0.66	0.78	0.01
% agreement between Coder 1 and Coder 2	84.2%	100%	97.4%	84.2%	98.7%	88.2%	94.7%	98.7%

As shown in the preceding table, very few FEs identified Task 8 (runs cables) as a separate task. Most likely, FEs thought that Task 8 was included under Task 2 (installs hardware and software). The poor identification of Task 8 was not surprising because it was not mentioned during a portion of the interview when most of the task list was collected. FEs would have had to add Task 8 later based on the rest of the interview, making it a harder task to identify. It's also apparent in the preceding table that the coders had relatively high levels of agreement identifying specific tasks. Noting that one of the coders changed between the pre-test and post-test, the average coder agreement was 88 percent in the pre-test and 93.5[2] percent in the post-test.

In addition to agreement between the two coders, Intraclass Correlation Coefficients (ICC) were also computed. The Intraclass Correlation Coefficient (ICC) generally ranges from 0 to 1.0 (although negative numbers are possible with some approaches), where numbers closer to 1.0 mean better reliability, and values near or less than zero suggest that agreement is attributable to chance alone. Many statisticians prefer that values be at least 0.6, and most often higher than 0.7, before claiming a good level of agreement.[3] In addition, two values of the ICC are possible: a single measure and an average measure. Since the single measure assumes only one rater, whereas the average assumes multiple raters, the single measure is reported in Table 6. In this case, a two-way random model was run with measures of consistent agreement because there were only two rating levels (correct or incorrect).

As shown in Table 6, and ignoring Task 8, it can be seen that the ICCs are generally moderate, with values above 0.6.

---

[2] Two experienced data collectors coded the tasks in the post-test, whereas in the pre-test one of the coders was inexperienced.

[3] Landis, J. R., Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33:159-174.

Table 6. Values of the Single Intraclass Correlation Coefficient (ICC) for Each Task

	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8
Pre-Assessment (N=97)								
Single ICC	0.48	0.614	0.665	0.733	0.398	0.532	0.911	0.918
Post- Assessment (N=76)								
Single ICC	0.41	1.000	0.910	0.691	0.927	0.738	0.844	0.000*

* The single ICC was zero for Task 8 because there was zero variance in the ratings (they were all 0).

4. Discussion

This calibration training session followed a series of three other calibration sessions that had been conducted to assess interviewer performance after initial training in a large field test. Calibration training is a tool that can be used to assess interviewer performance in addition to more traditional approaches such as reinterview, observation/monitoring, analysis of paradata, and quality reviews of questionnaires. However, calibration sessions focus on specific job skills and performance (Mockovak et al., 2015).

The key research objectives in this study were to determine how reliably and accurately Field Economists could code responses observed in a conversational interview and what impact four months of conversational interviewing experience had on coding performance.

Key findings were that interviewers could observe and record data at a high level of accuracy after initial training, and this skill increased significantly after four months of data collection, showing that interviewers' ability to code an observed conversational interview improved with job experience. The findings of this evaluation must also be considered in the context of extensive quality control procedures that have been put in place to introduce more consistency into data collection.

In addition, the evaluation approach provided specific information about performance on individual items and sections in the questionnaire, which could then become the focus of follow-up quality improvement efforts such as additional training. For example, although performance significantly improved on some sections (for example, physical demands and environmental conditions), it decreased on the job requirements section of the questionnaire. Earlier pretests had shown that the questions in this section were more difficult to code and a focused evaluation had explored the cause of these coding difficulties (Mockovak and Kaplan, 2015), but the results of the current study suggest that additional clarification and training may be needed.

The methodology used in this study was not ideal because interviewers did not actually have to conduct a conversational interview and only one interview could be observed in the pre and post sessions due to time and resource constraints. Therefore, the evaluation approach used should be viewed as an indirect measure of FE conversational interviewing ability because it focused on data recognition and recording skills, and not on the ability to use conversational interviewing to collect data. Data recognition and recording skills are important prerequisites for effective conversational interviewing, but additional data-collection skills are needed. Nonetheless, despite these drawbacks, this calibration training

session was a low cost, efficient approach for evaluating interviewer performance compared to other alternatives, and it provided timely and useful information.

In general, simple percent agreement provided the most intuitive measure of agreement and accuracy. In addition to determining the accuracy of responses (by comparing them to gold-standard answers), other measures of inter-rater reliability were investigated, but their results were less clear cut. For example, Cohen's kappa was not computed because of the very high levels of absolute agreement on many of the items, which can lead to low levels of estimated reliability. Instead, as an alternative, Cronbach's alpha was computed to produce a measure of internal consistency of the ORS instrument. This measure showed acceptable to excellent levels for the "Full Interview" and lower internal reliability for the post-course measure than for the pre-course measure, as well as for some sections of the instrument. However, as noted previously, these results can also be misleading. For example, although coefficient alpha for the physical demands section dropped between the pre and post measurements, the mean number of items correctly answered actually increased from 84.7 percent to 91 percent. Therefore, these additional assessments of inter-rater reliability should be interpreted cautiously. As noted by Thorndike:

> "*There is no single, universal, and absolute reliability coefficient for a test. Determination of reliability is as much a logical as a statistical problem.*"
> [Thorndike, 1951, pp. 570-57]

Also, as Stemler (2004) notes:

> "*the general practice of describing inter-rater reliability as a single, unified concept is at best imprecise, and at worst, potentially misleading.*" He goes on to note that "*Inter-rater reliability refers to the level of agreement between a particular set of judges on a particular instrument at a particular time. Thus, inter-rater reliability is a property of the testing situation, and not of the instrument itself.*"

Finally, the results of this evaluation say nothing about the validity of the data. As noted by Handel (2015), "… reliability refers to the reproducibility or repeatability of measurement values under similar conditions, while validity concerns the extent to which the measurement methods or quantities measured are appropriate or acceptable for the particular purpose(s) for which they will be used." The uses of ORS data and the precision required are topics beyond the scope of this paper.

Acknowledgements

References

Beatty, P. 1995. Understanding the Standardized/Non-Standardized Interviewing Controversy. Journal of Official Statistics, 11(2), 147-160.

Cohen J. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement; 20:37-46.

Cronbach, L.J. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika 16 (3): 297–334. doi:10.1007/bf02310555*

Ferguson, G. 2013. Testing the Collection of Occupational Requirements Data. Proceedings of the 2013 Joint Statistical Meetings, Montreal, Canada.

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

Gwet, K. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical and Statistical Psychology (2008), 61, 29–48.

Handel, M. J. 2015. Methodological Issues Related to the Occupational Requirements Survey**.** Internal Bureau of Labor Statistics report, February 26, 2015. Active URL: http://www.bls.gov/ncs/ors/handel_report_feb15.pdf

Harney, T. and Brown, K. 2015. Building Quality Assurance for the ORS. Proceedings of the 2015 Joint Statistical Meetings, Seattle, WA.

Meharenna, R. 2015. ORS Data Review Process. Proceedings of the 2015 Joint Statistical Meetings, Seattle, WA.

Maynard, D.W. and Schaeffer, N. C. 2002. "Standardization and Its Discontents." Pp. 3-46 in Maynard, D.W., Houtkoop-Steenstra, H., Schaeffer, N. C., van der Zouwen, J. (Eds.), *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview*, John Wiley & Sons, Inc.

Mishler, E.G. 1986. Research Interviewing: Context and Narrative. Harvard University Press.

Mockovak, W. 2010. Using an Action-Research Model to Move from Conversational to Hybrid Standardized Interviewing: A Case Study. In *JSM Proceedings,* Government Statistics Section. Alexandria, VA: American Statistical Association.

Mockovak, W. and Kaplan, R. 2015. Comparing Results from Telephone Reinterview with Unmoderated, Online Cognitive Interviewing. Poster presented at the 2015 annual meeting of the American Association for Public Opinion Research. Active URL: http://www.bls.gov/osmr/pdf/st150100.pdf

Mockovak, W.; Yu, A; and Earp, M. 2015. Using Calibration Training to Assess the Quality of Interviewer Performance. In *JSM Proceedings,* Government Statistics Section. Alexandria, VA: American Statistical Association.

Schober, M.F., Conrad, F.G. and Fricker, S.S. 2004. Misunderstanding standardized language. Applied Cognitive Psychology, 18, 169-188.

Suchman, L. and Jordan, B. 1990. Interactional Troubles in Face-to-Face Survey Interviews. Journal of the American Statistical Association, 85, 232-241.

Shrout, P.E. and Fleiss, J.L. 1979. Intraclass Correlations: Uses in Assessing Rater Reliability. Psychological Bulletin, Vol. 86, No. 2, 420-428.

Stemler, Steven. 2004. A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability. Practical Assessment, Research & Evaluation, ISSN 1531-7714.

Thorndike, R. L. 1951. Reliability. In E.F. Lindquist (Ed.), *Educational Measurement* (pp. 560-620). Washington, DC: American Council on Education.

Warrens, M. J. 2010. A Formal Proof of a Paradox Associated with Cohen's Kappa. *Journal of Classification*, Volume 27, Issue 3, 322-332.