MEASURING INTERVIEWER PERFORMANCE IN MOCK INTERVIEWS

Anitra Rustemeyer, Bureau of the Census

INTRODUCTION

One aspect of the work undertaken by the Census Bureau's Committee to Evaluate Initial Training of Interviewers was to develop a means of evaluating interviewer skill during the conduct of an interview. A paper published in Britain in the early 1950's found that only 12% of all errors made by interviewers could be detected by review of completed interview materials turned in by an interviewer; the remaining 88% were "invisible" during later review of completed materials in that they resulted from altering the scope of questions, probing and prompting errors, and incorrect recording of information.[1]

In A Technique for Measuring Interviewer Performance, Charles Cannell summarized the results of work done in recent years at the University of Michigan's Survey Research Center to systematically and objectively measure interviewer on-job performance. SRC's method differs from that used in Britain and in the Census Bureau in that it makes use of tape recordings of live interviews conducted in respondents' homes; whereas, the Census Bureau and the British studies used mock interviews in which staff members role-played as respondents.

OVERVIEW OF STUDY DESIGN

For our attempt to develop a test of interviewer performance we selected three probability samples of Census Bureau interviewers. The three samples represented:

1. "New" interviewers, who had just completed their initial home study and classroom training for the Current Population Survey (CPS) and had no field experience with CPS (N=72);
2. "EOT" (end-of-trng.) interviewers, who had completed all phases of initial CPS training (including on-the-job training) and had completed two or three field interviewing assignments (N=39); and,
3. "Expr"(experienced) interviewers, who had completed all initial training and had more than three months of field experience on CPS (N=114).

Although interviewers were sampled according to their levels of experience, those tested do not represent the interviewer work force as a whole. The proportion of new interviewers selected was greater than the proportion of experienced interviewers.

Each interviewer selected for the study was asked to conduct either three or four interviews with a staff member.[2] The persons who role-played as respondents followed a script so that each interviewer was tested on nearly identical situations. Only if the interviewer asked incorrect questions was the respondent allowed to deviate from the script. All interviews were tape recorded. Coders listened to all of the tapes and coded the quality of asking questions, probing, and introducing and closing the interviews. They also reviewed the completed questionnaires and coded them for consistency with the tape recording. Care was taken during coder training and quality control operations to assure that only one error was assigned for each mistake, and that interviewers not be penalized for mistakes of the "respondent" (tester). Independent check coding maximized uniformity of coder decisions.

For each interviewer included in the study, actions were evaluated for the following aspects of interviewing:

- --Asking questions
- --Probing for additional information
- --Recording answers
- --Filling transcription items
- --Introductions and closings
- --Accuracy of labor force classification.[3]

The first five aspects of interviewing were viewed in three ways: (1) Proportion of actions of each type that were judged to be correct actions; (2) a "score" for each interviewer which was calculated to give relatively more weight to actions considered by the analyst to have greater impact on the quality of data; and (3) proportion of the five types of errors that were of each type (without regard to impact on quality of data).

To judge accuracy of labor force classification, the questionnaires filled by the interviewers were subjected to a coding process that duplicated as closely as possible the Census Bureau's computerized labor force classification system.

RESULTS

While interviewers were sampled, the test they took did not sample situations they meet at work. The scripts were designed to be graded from easy to somewhat difficult. Analysis of consistency in interviewers' scores according to script will indicate how much test results described here are affected by the difficulties presented in the scripts. The following results, therefore, should be viewed as provisional:

Proportion of Correct Actions: Written vs. Verbal. As can be seen in Table A, interviewers were correct more often in their written work than in the verbal part of their job. Written entries were of acceptable quality 94-97% of the time, while the way in which questions were asked was judged to be acceptable 84-89% of the time, and the way in which interviewers probed for additional information was judged to be acceptable a little over 80% of the time.

Types of Errors and Frequency of Errors. Table B identifies the nature of the seven scores which were computed for each interviewer and shows for each of the three interviewer groups the mean, range, and standard deviation of the scores. Statistically significant differences were found between experienced interviewers and new interviewers for three of the scores: experienced interviewers were significantly better at filling transcription items and entering notes required by the answers given by respondents; also, the score summarizing the quality of all written work showed that experienced interviewers were significantly

better than inexperienced interviewers in that aspect of their work.

Nearly one half of all errors were related to how well interviewers asked questions. New interviewers made significantly more errors than did the experienced ones. These findings can be seen in Table C. It is also interesting to note (from Tables B and C) the extent of individual variation in number of errors made and test scores.

Relationships among Test Scores and Other Information about Interviewers. In order to examine differences among the S scores, correlation coefficients were computed (some are presented in Tables D1 and D2). All of the relationships among scores S1-S6 for experienced and new interviewers are positive and statistically significant.

Because the testing procedure used in this study is relatively expensive to administer, it was important to determine whether it provided new information about interviewers or whether it was largely a duplicate of some other measurement already in use and/or available at lower cost. The relatively large number of small and insignificant correlation coefficients shown in Table E support the conclusion that this test of interviewer performance does not merely provide a different way to approximate an existing measurement.

Visible vs. Invisible Errors. In order to compare our findings with the British study referenced above, their classification scheme was applied (see Table F). In Britain, the most common type of error was "failure to probe," while in our study the most common error of experienced interviewers was to "alter the scope of the question"; the most common type of error made by the inexperienced interviewers was what the British called "invisible recording errors."

The new interviewers made the highest proportion of visible errors (33% of the errors classified in Table F); at the end of their training period 18% of the interviewers' errors were visible; finally, experienced interviewers had only 9% of their errors in the "visible" category.[4] While the experienced interviewers made fewer errors than did new ones, and apparently had learned to avoid errors in the "visible" category, they were much more likely to alter the scope of the question. Visible errors are relatively cheap and easy to correct because they can be detected by means of an office review; therefore, it is disconcerting that 91% of the errors made by experienced interviewers were "invisible."

Quality of Labor Force Classification. Table G summarizes labor force classification results. It shows that 36% of the experienced interviewers made one or more errors that would have prevented labor force classification or resulted in the wrong classification. Sixty-seven percent of the inexperienced interviewers made such errors, while 61% of those with two or three months of experience made errors that prevented labor force classification or resulted in misclassification.

When considering the findings shown in Table G (as well as those shown throughout this report), it is important to bear in mind that for this study the performance of interviewers was judged in an artificial setting. Whether interviewers performed better or worse in this setting than in the field is a matter of surmise, but this test is predicated on an assumption that there is a relationship between the way an interviewer behaves in the field and performs on the test. As noted at the outset, the situations protrayed in the scripts used in our test were chosen to present interviewers with a variety of test situations; they should not be interpreted as a representative sample of situations encountered during the conduct of the Current Population Survey. Within this restriction on the generalizability of these findings, it is worthwhile to note that this study does provide evidence that errors made while administering surveys can result in misclassification of respondents. Whether the percentage of persons misclassified is 6%, as we found in the situations we contrived for our test, or whether it is some other percent cannot be determined by this study.

Although the procedures used in this study are reasonably expensive to follow, we are hopeful that something similar to the test and coding procedures we developed can be implemented at the Bureau as a way of giving each interviewer and his supervisors objective feedback on how well the interviewer is performing in several aspects of his job.

FOOTNOTES

1. Reported in Harris, Muriel, "Interview-Research: Paper VI, The Grading of Interviewers: An Examination of Visible and Concealed Interviewer Error as Revealed by the Grading Tests, and Some Suggestions for Future Grading Procedure," M52, Documents Used During the Selection and Training of Social Survey Interviewers and Selected Papers on Interviewers and Interviewing, The Social Survey Division, Central Office of Information, Great Britain, May 1952.

2. Interviewers in groups 1 and 3 were tested on four scripts (A,B,C,D); those in group 2 were tested on three scripts (B,D,E). The same person role-played as the household respondent for all mock interviews administered by an interviewer; persons who played respondent were regional office supervisors or professional staff from the Bureau's Statistical Research Division.

3. This was included as a measurement of the effect of interviewer errors on the quality of final data.

4. The difference between new and experienced interviewers in percentage of visible errors is statistically significant, i.e., in a difference of proportions test an approximate Z value of 4 was obtained. EOT was not tested with the other groups because its sample was so small.

TABLE A DISTRIBUTION OF SELECTED BEHAVIORS, BY LEVEL OF INTERVIEWER EXPERIENCE

Type of Behavior	Level of Interviewer Experience					
	Experienced		End-of-Trng.		New	
	%Acceptable	% Unaccep.	%Acceptable	% Unaccep.	%Acceptable	% Unaccep.
I. VERBAL BEHAVIOR IN INITIAL ASKING OF QUESTIONS						
Total verbal behaviors coded for how questions were asked	22,255=100%		5,466=100%		14,089=100%	
	89.4	10.6	84.3	15.7	86.8	13.4
A. Asked questions exactly as worded	62.0		60.2		62.2	
B. Changed only verb tense	2.9		2.3		1.8	
C. Made minor modification more than verb tense	19.6		15.6		18.1	
D. Made correct use of verification	4.9		6.2		4.7	
E. Rephrased question & changed meaning; read answer categories when not permitted; or made improper use of verification in lieu of asking question		5.2		3.6		3.9
F. Asked a question which should have been skipped		2.1		3.2		3.5
G. Failed to ask a question which should have been asked		3.3		8.9		6.0
II. VERBAL BEHAVIOR AFTER INITIAL ASKING OF QUESTION (PROBING)						
Total number of probing behaviors coded	5,289=100%		1,202=100%		2,714=100%	
	80.7	19.4	86.0	14.1	80.3	19.6
A. Repeated questions correctly	5.6		7.6		7.1	
B. Made up non-directive probe	38.4		50.2		45.4	
C. Correctly repeated or summarized respondent	35.4		28.0		25.8	
D. Correctly confirmed frame of reference	1.3		.2		2.0	
E. Failed to probe when necessary		6.7		8.5		8.3
F. Probed directively		8.6		3.2		6.8
G. Incorrectly verified respondent's answer		2.2		1.2		2.2
H. Added to question incorrectly; repeated question or part of it incorrectly; confirmed incorrect frame of reference; or probed unnecessarily		1.9		1.2		2.3
III. WRITTEN BEHAVIORS	N=49,105		N=11,806		N=31,174	
	97.1	2.9	93.9	6.1	94.4	5.6
A. Number of items judged for quality of recording answers to survey questions	24,350=100%		6,069=100%		15,598=100%	
	95.5	4.5	91.2	8.8	94.9	5.1
1. Made correct entry	90.6		82.5		86.1	
2. Made entry consistent with verbal; but incorrect info was obtained due to interview verbal error	4.6		8.4		8.7	
3. Made entry consistent with verbal; but incorrect info was obtained due to respondent verbal error	.3		.3		.1	
4. Recorded info correctly; but it was not obtained in interview (usually means Ir guessed)		.5		1.7		.5
5. Entry or lack of entry was inconsistent with verbal (not used when (4) above applies)		3.9		6.2		4.5
6. Entry was in incorrect location; but intent was clear		.1		.3		.1
7. Omitted entry correctly		NA		.6		NA
B. Number of items judged for quality of filling transcription items	24,481=100%		5,576=100%		15,411=100%	
	98.8	1.3	97.8	2.2	94.2	5.8
1. Made a required entry correctly	98.8		97.8		94.2	
2. Made a required entry incorrectly		.8		.9		1.6
3. Failed to make a required entry		.5		1.3		4.2
C. Number of items judged for quality of entering required notes[1]	274=100%		161=100%		165=100%	
	88.7	11.3	59.0[2]	41.0[2]	70.9	29.1
1. Required note present and correct	88.7		59.0		70.9	
2. Required note not present		4.7		36.6		21.2
3. Required note present but not correct		6.6		4.4		7.9

TABLE A FOOTNOTES

[1]For the most part these were instances in which the interviewer marked an answer category which contained the instruction: "Specify" or "Specify in Notes".

[2]Comparison of EOT interviewers with the other two groups should not be made in this section as three-fourths of the "unacceptable" behaviors occurred in Script E, which was not used to test new and experienced interviewers.

TABLE B DESCRIPTION OF WEIGHTED SCORES ACHIEVED ON THE MIP TEST BY LEVEL OF INTERVIEWER EXPERIENCE

Type of Behavior Scored	Test Score[1]									t-Values for Expr. and New[2]
	Mean			Standard Deviation			Range			
	Expr	EOT	New	Expr	EOT	New	Expr	EOT	New	
S1. Asking Questions	574.9	473.2	524.0	155.9	137.9	149.4	178-905	178-811	164-860	1.931
S2. Probing	639.9	727.6	652.9	126.3	111.2	129.8	288-947	481-964	383-957	.033
S3. Written Entries	472.9	315.5	429.2	119.4	93.6	127.7	242-1000	119-540	183-982	2.169
S4. Recording Answers	434.4	286.1	407.8	121.4	95.4	129.3	214-1000	98-505	176-1000	1.288
S5. Filling Transcription Items	855.7	799.2	664.9	126.0	164.2	212.3	315-1000	364-1000	54-1000	7.089
S6. Entering Required Notes	790.1	353.5	524.1	369.7	234.8	440.1	000-1000[3]	000-1000[3]	000-1000[3]	4.592
S7. Introductions to and Closing of Interviews	399.5	361.6	433.9	201.7	128.9	147.3	141-1000	118-688	208-1000	-1.532

[1]The scores shown were computed on a scale of 0 to 1000, where 1000 is the best score possible. In forming the scores, some behaviors were given relatively more weight than others in order to reflect the opinion that all behaviors are not of equal importance. In computing the S scores, the weights were applied to the frequency counts and then the weighted count of acceptable behaviors was divided by the weighted count of all behaviors.

[2]In the test used, a positive t-Value means that the experienced interviewers were higher; a negative value means that new interviewers were higher. A value greater than 2 or less than -2 is statistically significant at the 5% level. EOT results were not tested with the other groups because of small sample size.

[3]This proportion is meaningless because it is often based on only 1 or 2 behaviors.

TABLE C SOME STATISTICS ABOUT THE NUMBER AND TYPE OF ERRORS MADE

Type of Error	Number of Errors Made									t-Values for Expr. and New[1]
	Mean			Stnd. Deviation			Range			
	Expr.	EOT	New	Expr.	EOT	New	Expr.	EOT	New	
Total Number of Errors Made	50.9	50.1	64.7	18.9	18.1	34.5	12-114	22-106	17-219	-2.722
E1. Asking	24.4	23.8	29.1	13.3	11.7	14.3	4-68	6-59	5-84	-1.823
E2. Probing	10.4	5.6	8.7	4.1	2.2	4.0	2-26	1-13	2-19	2.843
E3. Recording Answers	9.9	14.5	11.7	4.2	5.7	5.2	0-24	6-38	0-30	-2.307
E4. Transcription	2.6	3.1	12.4	3.0	4.0	22.3	0-21	0-20	0-138	-4.002
E5. All-Other Errors	3.6	3.1	2.7	2.1	1.3	1.6	0-11	1-7	1-9	4.169

[1]In the test used a positive t-Value means that the experienced interviewers were higher; a negative value means that new interviewers were higher. A value greater than 2 or less than -2 is statistically significant at the 5% level.

TABLE D1 CORRELATIONS AMONG TEST SCORES AND PERCENT OF ITEMS WITH PROBES FOR EXPERIENCED INTERVIEWERS

	S2	S3	S4	S5	S6	S7	Percent w/Probes
S1	.3750*	.5124*	.4953*	.2774*	.2613*	.0791	-.0561
S2		.3963*	.3858*	.2397*	.3001*	.1372*	.3791*
S3			.9960*	.3319*	.3358*	.0013	.1368
S4				.2624*	.3094*	-.0021	.1233
S5					.2413*	.0779	.1599
S6						.0505	.0483
S7							.0399

TABLE D2 CORRELATIONS AMONG TEST SCORES AND PERCENT OF ITEMS WITH PROBES FOR NEW INTERVIEWERS

	S2	S3	S4	S5	S6	S7	Percent w/Probes
S1	.2904*	.4934*	.4411*	.6346*	.3788*	.1819*	.1303
S2		.2447*	.2093*	.3570*	.3227*	-.0525	.2831*
S3			.9796*	.5459*	.2439*	.1248*	.0374
S4				.3938*	.2155*	.1235*	.0495
S5					.3083*	.0914	.0353
S6						-.1673*	.1808
S7							.1193

*Statistically significant at 5% level by Fisher's Z-statistic.

S1 -- Asking questions
S2 -- Probing
S3 -- Written entries (combination of S4,S5, & S6)
S4 -- Recording answers
S5 -- Filling transcription items
S6 -- Entering required notes
S7 -- Introductions and closings
Percent w/Probes--Percent of items on which probing was done

TABLE E CORRELATIONS (FOR EXPERIENCED INTERVIEWERS) BETWEEN TEST SCORES AND OTHER INFORMATION ABOUT INTERVIEWERS

	S1	S2	S3	S4	S5	S6	S7	D2	D3	D4	D5
D1	.0768	-.1536	.1614	.1654	-.0599	.1132	.0225	.0405	.0700	.0521	.0186
D2	-.3218*	-.2317	-.1672*	-.1692	-.0344	.0721	-.1725*		.0221	.2634*	-.0816
D3	-.2811*	-.2727	-.1264	-.1170	-.1455	-.3412	-.2193			.1217	.2490*
D4	.0469	.0563	.0581	.0560	.0071	.0514	.0605				.0056
D5	-.3073*	-.1904	-.0916	-.0874	-.0992	-.2810*	-.0445				

*Statistically significant at 5% level by Fisher's Z-test.

D1 -- Education
D2 -- Age
D3 -- Error Rate at time of test
D4 -- Number of minutes used to complete test
D5 -- Non-interview (of eligible households) Rate: weighted average for 3 months prior to test
S1 -- Asking questions
S2 -- Probing
S3 -- Written entries
S4 -- Recording answers
S5 -- Filling transcription items
S6 -- Entering required notes
S7 -- Introductions and closings

TABLE F DISTRIBUTION OF INTERVIEWER ERRORS BY TYPE AND LEVEL OF INTERVIEWER EXPERIENCE; COMPARED TO A BRITISH STUDY

TYPE OF ERROR	PERCENTAGE OF TOTAL ERRORS				
		Interviewers in Mock Interview Project (MIP)			
	Britain[2] (N=56)	Total (N=225)	Experienced (N=114)	End-of-Trng. (N=39)	New (N=72)
Total Errors[1]	1288=100%	7750=100%	3762=100%	1080=100%	2908=100%
Invisible Errors					
1. Failure to probe for additional information i.e. to find out if informant has anything further to add; and failure to probe sufficiently to establish criteria laid down in instructions or definitions or to clarify ambiguous answers.	34	9	9	9	8
2. Overprobing after it has become clear that informant has nothing further to add; or failure to recognize that they have all the information they require to classify.	7	1	1	1	1
3. Altering the scope of the question.	17	33	42	22	26
4. Prompting errors--failure to prompt when instructed, omission of items on prompt list, reading prompt list before all spontaneous information has been obtained.	1	8	9	7	6
5. Invisible recording errors. Any recording errors which could be discerned at the coding stage have been excluded from this category & dealt with the Category 6.	29	30	29	43	27
Visible Errors					
6. All errors discernible at the coding stage i.e. anything that appears to be an error in the light of other evidence on the schedule, omissions or inadequate information, items written in the wrong place and answers put under "others" when they fit a precode.	12	19	9	18	33

1. For the purpose of comparing the MIP results with the British study cited in Footnote 2, "total errors" is defined as it was in the British study. In the MIP the following additional types of errors were classified: incorrect selection of questions to be asked, asking questions out-of-order, and incorrect introduction to and closing of interviews.

2. Reported in Harris, Muriel, "Interviewer-Research: Paper VI, The Grading of Interviewers: An Examination of Visible and Concealed Interviewer Error as Revealed by the Grading Tests, and Some Suggestions for Future Grading Procedure," M.52, Documents Used During the Selection and Training of Social Survey Interviewers and Selected Papers on Interviewers and Interviewing, The Social Survey Division, Central Office of Information, Great Britain, May, 1952.

TABLE G SUMMARY OF LABOR FORCE CLASSIFICATION ERRORS OF INTERVIEWERS TESTED, BY DURATION OF EMPLOYMENT ON CPS

	All Interviewers (N=225)	Experienced (N=114)	End-of-Training (N=72)	New (N=39)
1. Percent of interviewers who made one or more errors affecting ESR classification	49.1	36.0	61.1	66.7
2. Number of persons portrayed in test scripts*	3008	1610	1008	390
3. Number of those on line 5 who were unclassifiable or misclassified	180	52	88	40
4. Line 3 as a percent of line 2	6.0%	3.2%	8.7%	10.3%
5. Mean number of unclassifiable and misclassified persons per interviewer	.80	.45	1.22	1.03

*This is the number of persons portrayed in each script, multiplied by the number of interviewers who were tested with the script.