A COMPARISON OF TWO BEHAVIOR CODING SYSTEMS FOR PRETESTING QUESTIONNAIRES

W. Sherman Edwards, Vasudha Narayanan, and Stephanie Fry, Westat; Joseph A. Catania and Lance M. Pollack, UCSF Health Survey Research Unit

Introduction

Coding of interviewer and respondent behavior is well established as a tool for evaluating survey questionnaires (as noted by Schaeffer, 1991, e.g.), particularly during pretests. Perhaps the most widely used behavior coding system for questionnaire evaluation was developed at the University of Michigan. As described in Oksenberg et al (1991), coders listen to interviews and assign codes to the interviewer's reading of a question (read exactly as worded, read with slight change, read with major change) and to the respondent's subsequent behavior (interruption with answer, request for clarification, adequate answer, qualified answer, inadequate answer, don't know, refusal to answer). For a given question in a given interview, a coder assigns one interviewer behavior code and one or more respondent code, and the codes are tallied for each question across coded interviews. Each code other than "read exactly as worded" or "adequate answer" may indicate a problem with the question.

Oksenberg et al used behavioral coding, special probes designed into the interview, and coder and interviewer debriefing to evaluate the pretest questionnaire in their experimental study. They conclude that behavior coding is reliable and efficient, that it can be implemented either live or with tape recorded interviews, and that a simpler coding scheme than the one described in the 1991 paper would be quite effective for pretesting questions. Interpretation of behavioral coding results is both a quantitative and qualitative exercise. Oksenberg et al note that tallying codes does not necessarily identify the reason for apparent problems. Schaeffer also points out that applying statistical tests to tallies of behavior codes may not be appropriate because of the clustering of behaviors by interviewer and respondent.

Westat has sporadically used behavior coding to evaluate pretest questionnaires, and has not yet centralized the process, so coders are typically inexperienced. Westat researchers have reported (personal communications) that the system described by Oksenberg et al is too complex to institute live, particularly when the coders are telephone center team leaders with other responsibilities.

In one recent pretest of a telephone survey, Westat reduced the behavior coding scheme to a single code – whether or not the respondent gave a response to the question that the interviewer could record without probing or offering clarification. Essentially, this simplified system combines all of the Oksenberg et al respondent problem codes except "interrupts with answer" and, perhaps, "qualified answer." All behavioral coding was done live. Results of this coding were combined with qualitative monitoring and interviewer debriefing to inform final revisions to a telephone survey questionnaire.

After this experience, we wondered how much more information would have been provided by a more complete set of codes, how comparable the results would have been in assigning codes, and whether live coding differed from coding of tape recorded interviews. This paper describes such a comparison made on a pretest of the Urban Men's Health Study 3.

The UMHS-3 is the third in a series of random-digit-dial surveys of men who have sex with men (MSM) in San Francisco, conducted by researchers at the University of California at San Francisco (UCSF). The UMHS-3 will be used to compare the extent of risky sexual behavior of MSM with and without a history of childhood sexual abuse. Westat is collecting the survey data. Part of Westat's responsibility was to conduct a pretest, using essentially the evaluation techniques described by Oksenberg et al.

Methods

UCSF identified some 61 questions in the draft UMHS-3 interview to target for behavioral coding. Westat conducted 41 pretest interviews, using three interviewers. All interviews were tape recorded. We trained seven behavioral coders for four hours in the Oksenberg et al system, including research assistants, telephone center operations managers, and telephone center team leaders. The team leaders were not subsequently used for the actual coding. All 41 interviews were coded from tape; seven were also coded live. Senior researchers also re-coded ten of the interviews from tape.

Based on the results of the behavior coding, responses to the special probes for 7 of the targeted questions, and an interviewer debriefing, UCSF and Westat revised the draft questionnaire.

Subsequently, the same coders re-coded 40 of the 41 interviews from tape, using the simplified coding scheme described earlier. Training for this round of coding took an hour and a half. Again, senior researchers re-coded ten of the interviews.

With two exceptions, none of the coders or re-coders listened to the same interview they had in the first round.

Results – Incidence of codes and inter-coder reliability

Table 1 compares the results of the first round of coding with those reported by Oksenberg et al. The first set of columns compares the mean incidence of each problem code. Although one would not necessarily expect close agreement between these two studies using different questionnaires, different interviewers, and a different respondent population, most of the incidence levels are strikingly similar. For both "inadequate answer" and "qualified answer," though, the Oksenberg et al incidences are more than three times those of the UMHS-3.

The second and third sets of columns compare the proportion of question items meeting different "problem thresholds" — either 25 percent or more or ten percent or more of times a question was asked the problem code was applied. Problems with reading questions were more concentrated in the UMHS-3 than the Oksenberg et al study, as were requests for clarification. Again, inadequate responses were much more frequently a problem in the Oksenberg et al study.

The next set of columns compares the "reliability" of the coding with the Kappa statistic (described in Maclure and Willett (1987), e.g.). The Oksenberg et al coders agreed consistently more often than the Westat coders, although Kappas for the Westat coders were in the "very good" range for three of the six categories. There was substantially less agreement among Westat coders on "inadequate answer," "qualified answer," and "slight change." Review of 78 questions where the initial coder and re-coder disagreed on assignment of a response problem code revealed that in only 7 cases did the coders disagree on *which* code should be assigned; in all others the disagreement was on *whether* a problem code should be assigned or not.

Finally, the last column of Table 1 presents Kappa statistics for agreement between the UMHS-3 live and taped coding. The first four categories are reasonably comparable between modes, but the reliability falls off even lower for the live versus taped comparison on "inadequate answer" and "qualified answer."

Table 1. Comparison of Oksenberg et al (1991) and UMHS3 behavior coding results

	Mean Incidence of Problem		Problem Indicator 25% Level		Problem Indicator 10% Level		Inter-coder Reliability: Kappa Values		
	Oksenberg	UMHS3	Oksenberg	UMHS3	Oksenberg	UMHS3	Oksenberg	UMHS3 (recode from tape)	UMHS3 (tape vs. live)
Question Asking:									
Slight change	12%	13%	17%	23%	50%	44%	0.73	0.28	0.38
Major change	4%	3%	0%	5%	10%	15%	0.72	0.67	0.45
Responses:									
Interrupts	4%	4%	3%	2%	12%	10%	0.90	0.67	0.59
Requests clarification	10%	9%	10%	21%	50%	30%	0.93	0.70	0.66
Qualified answer	7%	2%	8%	2%	20%	4%	0.56	0.15	0
Inadequate answer	24%	7%	35%	0%	72%	28%	0.85	0.49	0.12
Don't Know	4%	1%	3%	2%	12%	4%	0.86	N/A	N/A
Refused	0%	0%	0%	0%	0%	0%	N/A	N/A	N/A

Oksenberg et al (1991) -- 60 questions, 60 interviews, Kappa based on 19 Interviews recoded
UMHS3 -- 61 questions, 41 interviews (many skips), Kappa based on 10 interviews recoded

Table 2 compares agreement between the first and second rounds of UMHS-3 coding[1], with the assumption that the Wave 2 problem code encompasses the four listed Wave 1 categories (there were no item refusals). For the second round of coding, using the simplified system, 16 percent (291) of items were identified as having a response problem, as compared with 14 percent (255) in the first round. Of the 255 first round problems, 160 (63 percent) were also identified as problems in the second round. The most problematic category was "inadequate answer," with only 20 percent of first-round problems flagged in the second round.

Using the "any problem" totals for the first round, overall agreement between the two rounds, as measured by Kappa, was 0.38. Review of a sample of 44 disagreements against the tape recordings revealed

[1] This comparison uses the initial coding from tape for all tallies.

Table 2. Agreement between Round 1 and Round 2 UMHS-3 coding

	Round 2 Code:		
	Problem	Not a problem	Percent Agreement
Round 1 Code:			
Requests clarification	102	15	87%
Qualified answer	12	10	55%
Inadequate answer	16	63	20%
Don't Know	7	3	70%
Multiple problem	23	4	85%
No response problem	131	1480	92%
Total items with problem	160	95	63%
No response problem	131	1480	92%

Kappa for Problem/Not	0.38

that only 5 were "false positives," or instances where one coder assigned a problem code in error. In one of these "false positive" problems, the respondent gave a codeable answer, but other comments indicated that he had clearly misunderstood the question. Many of the other disagreements were fairly subtle problems. For example, several respondents answered a long series of questions about their emotions, for which the response categories were "never," "once a month," "once a week," etc., with "no." This is clearly not a codeable response (inadequate), but the sheer repetition of the response gave it credence for some coders. This problem was the source of most of the 63 "inadequate response" codes from the first round that were not marked as a problem in the second, as well as of many of the second round problems not marked in the first. As another example, some respondents would questioningly repeat part of the question or the answer categories, get confirmation from the interviewer, and then provide a codeable response. Some coders did not consider this behavior "requesting clarification."

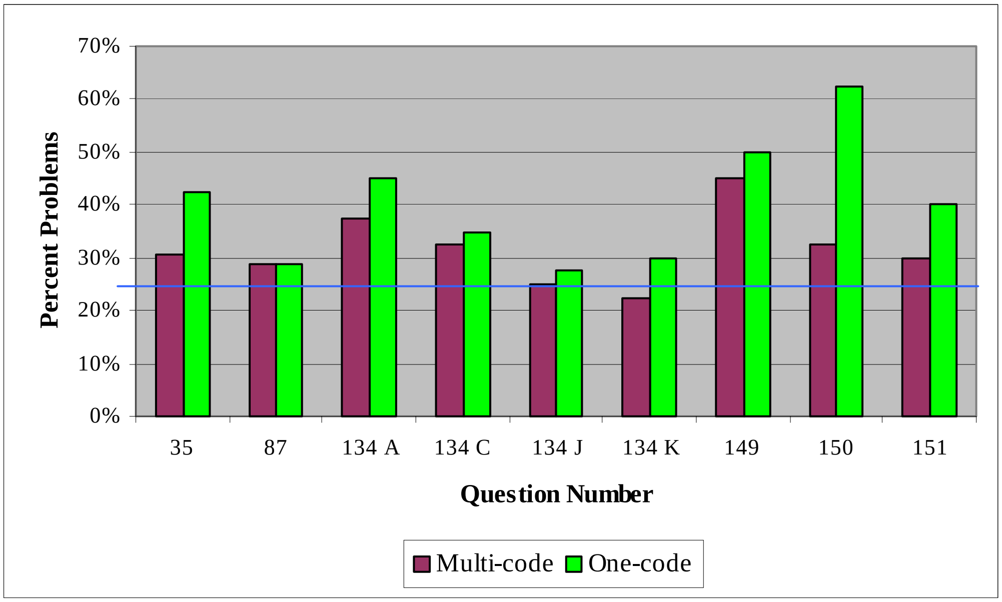
Kappa for the second round as measured by the re-coding, was 0.84, substantially higher than for any of the categories of the first round of UMHS-3 coding. However, the high level of disagreement between the first and second rounds, and the nature of those disagreements, indicates that the "true" level of problems was under-identified in both rounds.

Results – Identifying questionnaire problems

Chart 1 shows all of the question items among the 39 items coded at least nine times that were identified as having response problems at the 25 percent level (the horizontal line in Chart 1) either in Round 1 (multi-code) or Round 2 (one-code). All of these questions but one (Q. 134K) were identified as problems at the 25 percent threshold by both coding schemes. Also for all questions but one (Q. 87), a higher proportion of cases was identified as problems in Round 2 than in Round 1.

In Round 1, seven items were identified as "slight change" question-reading problems at the 25 percent level. Only one of these items was also among the response problem group in Chart 1. Three items were coded as "major change" problems, all among the "slight change" problem group. One item was a problem at the 25 percent level with "respondent interrupting with answer"; it was also a "major change" and a "slight change" problem. Thus, the question-reading codes definitely added to the identification of problem questions, but there was overlap among the three different problem codes in what questions were identified.

Chart 1 – Comparison of Round 1 (multi-code) and Round 2 (single-code) systems in identifying questionnaire items with response problems



Discussion

In this application of behavior coding to evaluate a pretest questionnaire, the coding provided useful information, in combination with information from observations and debriefings, for revising the questionnaire. The application of the multi-code system did not yield as high a proportion of problems as reported by Oksenberg et al (there is no particular reason one should expect it to), and the Westat inter-coder reliability was somewhat lower than reported by Oksenberg et al. More extensive training of the coders based on the experience with this study would likely increase both the number of problems identified and the inter-rater reliability in a subsequent application of the system. However, neither the project nor coding staff may have a similar assignment in the near future.

Live coding produced somewhat fewer problem codes and lower inter-rater reliability than coding from tape. At least among coders with the level of training and experience of those used in this study we do not recommend relying solely on live coding using the multi-code system.

The single-code system produced a higher incidence of response problem codes than the multi-code system, and yielded higher inter-rater reliability. Reliability between the two systems was relatively low (Kappa = 0.38), but investigation of coding differences did not indicate any systematic difference in what should be coded as a problem. As one would expect, the single-code system seems to be easier to learn and apply consistently than the multi-code system. The multi-code system does provide more detail about the kinds of respondent behavior identified, but not sufficient detail to diagnose and correct the specific problems with the questionnaire items without other, qualitative input.

The two systems were very similar in identifying questionnaire items with response problems at the 25 percent incidence level. The multi-code system also identified questionnaire items as having reading problems, which largely did not overlap with the response problem items. However, two of the three reading problem codes overlapped completely with the third code. Thus, it appears that the multi-code system will identify more problem question items than the single-code system. It may be possible to add a second code to the single-code system, perhaps combining the three problem codes “small change,” “major change,” and “respondent interrupts with answer.”

This paper has reported on a case study in the application of behavior coding for questionnaire pretesting. The sample sizes were relatively small, and a small group of coders was used. The respondent population was urban men who have sex with men, who tend to be more highly educated and more affluent than the general population. Many of the questions would be considered sensitive, as they deal with sexual behavior and HIV/AIDS (although this population tends to be relatively open about discussing these issues). Of course, every questionnaire has its own idiosyncrasies and would be expected to have a unique pattern of reading and response errors. While for these reasons generalizing from this study is problematic, it does appear that a simplified behavior coding system may be quite useful in pretesting questionnaires, particularly for organizations or groups within organizations that do not regularly do such coding.

We anticipate developing and using a two-code system, one code for response problem and one code for reading problem, and hope to be able to compare this system to the Oksenberg et al multi-code system. While we have used the one-code system with live monitoring of telephone interviews, we did not evaluate its effectiveness in this experiment. We anticipate extending future work to comparing live versus taped coding using a simplified coding system.

References

Maclure, M., and Willett, W. (1987), Misrepresentation and Misuse of the Kappa Statistic, American Journal of Epidemiology, Vol. 126 No. 2, pp. 161-9.

Oksenberg, L., Cannel, C., and Kalton, G. (1991), New Strategies of Pretesting Survey Questions, Journal of Official Statistics, Vol. 7, No. 3, pp. 349-366.

Schaeffer, N. (1991), Conversation with a Purpose—or Conversation? Interaction in the Standardized Interview, in Biemer et al, eds., Measurement Errors in Surveys, John Wiley & Sons, Inc., New York.