PREDICTING TEST-RETEST RELIABILITY FROM BEHAVIOR CODING

Jennifer C. Hess, SRD/CSMR, Bureau of the Census, Rm. 3125-4, Washington, DC 20233
Eleanor Singer, Survey Research Center, University of Michigan

Key Words: Behavior coding, Reinterview, Question reliability

INTRODUCTION

In attempting to move questionnaire design from art to science, researchers use different evaluation techniques to help determine how well questions are working. Techniques such as behavior coding, respondent debriefing, interviewer debriefing, cognitive interviewing, and nonresponse analysis all provide information to help the questionnaire designer assess whether respondents understand questions as intended and whether they are able to provide adequate answers to them. In 1994, Presser and Blair evaluated some of these methods, concluding that behavior coding provided more reliable diagnoses of question difficulties than conventional pretests involving a small number of interviewers followed by an interviewer debriefing.

However, with the possible exception of some types of respondent debriefing questions, these techniques do not actually measure question reliability. Reliability data, such as those that could be obtained in a test-retest experiment (reinterview), are rarely collected as part of pretest activities because they are time-consuming, labor intensive and very costly to collect. Of course, the goal of good questionnaire design is to produce reliable and valid information, not simply questions that are easy for respondents to answer. But it is assumed that questions that pass the screen of the questionnaire evaluation techniques described above are also more likely to produce data that are reliable and valid.

How well do question evaluation techniques in fact predict reliability and validity? Data reported by Belli and Lepkowski (1995) suggest that interviewer behaviors have little predictive value for response accuracy, though respondent behaviors are somewhat more predictive of response accuracy. Recently, the U.S. Department of Agriculture's Food and Consumer Service fielded a new survey, designed to measure the subjective experience of hunger in the United States. This survey provided an opportunity to examine how well some traditional question evaluation techniques predict test-retest reliability. The Census Bureau was asked to help develop the questionnaire, using some of the evaluation methods listed above. In addition, a reinterview was conducted with a sample of households following the survey. In this paper, we use behavior coding data to predict how reliably questions are answered, as measured by an index of inconsistency developed by the Census Bureau.

METHODS

Sample

The Food Security Supplement to the Current Population Survey (CPS) was conducted from April 16-25, 1995 on a nationally representative sample of approximately 54,000 interviewed households. Respondents were asked both the CPS labor force questions and the Food Security Supplement questions. The response rate for the CPS was 92.9 percent and for the supplement was 85.4 percent. Approximately 90 percent of the cases were conducted in the field using computer assisted personal interviewing (includes both personal visit interviews and telephone interviews from field representatives' homes) and 10 percent were conducted at the Census Bureau's centralized telephone facilities using computer assisted telephone interviewing.

Approximately 34 percent of the households in the sample were "low income," which, for the purposes of this study, is defined as at or below 185 percent of the poverty level.[1] Three-quarters of the sample households were urban and one-quarter rural. Approximately 85 percent of the households were White, 10 percent were Black, and 6 percent were Hispanic (could be of any race).[2]

The questionnaire included five different sections: food expenditures, program participation, food sufficiency, coping mechanisms and food scarcity, and concern about food sufficiency.[3] Food expenditures were asked of all households. These questions collect information on the actual amount the household spent for food last week and the usual amount the household spends on food per week. The program participation section asks about food stamp recipiency and participation in other government and private programs that provide food, such as the school lunch program and WIC. The food sufficiency section contains questions used to assess whether respondents clearly have enough to eat or whether there are times when their resources are strained and they have difficulty providing themselves or their families with a nutritionally adequate diet. These questions are used to screen respondents either into or out of the remainder of the questionnaire. The coping mechanism and food scarcity section measures the extent of food insecurity in the household as do the questions in the section on concern about food sufficiency.

Behavior Coding

Behavior coding is the systematic coding of the interactions between an interviewer and a respondent

(Cannell, Lawson, and Hausser, 1975; Cannell et al., 1989). Interviewers at the Census Bureau's Hagerstown and Tucson Telephone Centers tape recorded a total of 147 cases of which 136 were subsequently behavior coded. (Eleven cases were not used because permission to record the interview was not on the tape.) We used a quota sample for behavior coding, not a random sample. The telephone centers were instructed to tape record interviews with the first 75 low income households.

We coded the first exchange between the interviewer and the respondent for each question. Coders assigned one interviewer code and up to two respondent codes per question. (Two respondent codes were most often assigned when the respondent interrupts the question reading to provide an answer. Thus, one of the codes is a "break-in" and the other may be any of the remaining respondent codes.) Four experienced coders from the Hagerstown Telephone Center behavior coded the tapes.[4]

To assess coder reliability, each coder was asked to complete the same five cases (in addition to the regular workload). The coders averaged 87 percent agreement on interviewer codes, 92 percent agreement on at least one of the two respondent codes, and 83 percent agreement on both respondent codes. The kappa statistics, which take into account the probability that two coders will agree on a code by chance, ranged from .68 to .80 for between coder agreement on interviewer codes, .74 to .93 on at least one of the two respondent codes, and .55 to .84 on both respondent codes. Kappa values above .75 represent excellent agreement and values from .40 to .75 represent fair to good agreement beyond chance (Fleiss, 1981). Thus, our statistics indicate fair to excellent agreement between coders.

An evaluation of the supplement questionnaire based on behavior coding data indicated that the food expenditures section caused the most problems of any section (see Table 1). Eighty-three percent (N=18 questions) of the questions in this section were flagged as problematic by behavior coding. Approximately 60 percent of the questions in the food sufficiency section (N=10 questions) and the concern about food sufficiency section (N=6 questions) were problematic. The remaining two sections, the program participation section and coping mechanisms and food scarcity section, caused fewer problems. Twenty percent of the questions in the program participation section (N=10 questions) and 28 percent of the questions in the coping mechanisms and food scarcity section (N=36 questions) were problematic. However, 15 of the 36 questions in the latter had less than 7 responses. When these cases are excluded, the percentage of problematic cases in this section drops to 10 percent. (Results are for both categorical and continuous variables.)

Table 1. Percentage of Problematic Supplement Questions By Section

Section	Total number of questions in section	Percent problematic questions
Food expenditures	18	83 %
Program participation	10	20 %
Food sufficiency	10	60 %
Coping mechanisms and food scarcity	36 21	28 % 10 % (excluding questions with less than 7 cases)
Concern about food sufficiency	6	67 %

Reinterview

The Food Security Supplement reinterview was conducted from April 17-29, 1995 by CPS supervisors, senior field representatives, and interviewers. Approximately 90 percent of the reinterviews were conducted within 7 days of the original interview, but in some cases, there was up to a 10 day lag.[5] The reinterview was conducted on a nationally representative sample of 1,827 with a response rate of 63.6 percent (1,162 completed interviews). The reinterview was conducted with the same respondent who had answered the original survey. The sample was split between households with family incomes at or below 185 percent of the poverty level and those with family incomes above 185 percent of the poverty level; 929 reinterviews were conducted with the former group and 233 with the latter. This sample was drawn in order to test two important features of the questionnaire: 1) the reliability of the screening questions that determined whether a respondent was asked the remaining questions that measure degree of food insecurity, and 2) the reliability of the questions on food insecurity. Because of cost constraints, most reinterviews were conducted by telephone.[6]

The major objective of the reinterview was to measure response variance, that is, to determine the degree of inconsistency between the original survey answer and the reinterview answer. The reinterview data contain several measures of response variance. We will use the index of inconsistency in this paper. This is a relative measure of response variance that estimates the ratio of response variance to total variance for each question. In general, an index of less than 20 indicates that response variance is low; an index between 20 and 50 indicates that response variance is moderate; and one over 50 indicates that response variance is high (McGuinness, forthcoming).[7]

Table 2 shows the mean and median index of inconsistency by section of the questionnaire for categorical variables.

Table 2. Mean and Median Index of Inconsistency for Each Section of the Questionnaire

Section	Mean	Median
Food expenditures	52	52
Program participation	25	19
Food sufficiency	46	47
Coping mechanisms and food scarcity	44	44
Concern about food sufficiency	53	52

In general, these data indicate that four of the five sections of the questionnaire are producing moderately to highly unreliable data, with the notable exception of the program participation section.

RESULTS

Behavior coding guidelines generally state that a question is considered problematic if less than 85 percent of the time interviewers read questions exactly as written or with only slight changes that do not affect question meaning, or if less than 85 percent of respondents give adequate or qualified answers to the question (Oksenberg, et al., 1991). Our analysis is limited to questions with a minimum of 7 cases in the behavior coding data.

We compare the results of behavior coding to those of the reinterview data at the question level. That is, we compare the diagnostic utility of behavior coding in predicting which questions will yield reliable data on reinterview. We do not have matching datasets at the level of the individual respondent, since the samples for behavior coding and for reinterview were drawn independently.

The questionnaire contained 75 questions, plus one split ballot item. There were 55 categorical questions of the "mark one answer" type, 20 continuous questions, and one question that was a "mark all that apply" type. This question had 5 possible responses and is treated as five separate questions in this analysis.

We were unable to use all questions in our analysis for two reasons. First, 3 questions were excluded because they had less than seven cases in the behavior coding data, 16 were excluded because of an unreliable index of inconsistency, and 15 were excluded because of both reasons. In most cases, the index was unreliable because the characteristic of interest is rare in the population and too few respondents were reinterviewed to provide reliable estimates. Thus, 46 questions were available for analysis. Second, because the index of inconsistency is calculated differently for categorical and continuous variables and the small number (N=9) of continuous variables made it impossible to carry out separate analyses for them, we decided to restrict the analysis to categorical variables.[8] The analysis in this paper is, therefore, restricted to the 37 categorical variables for which we have reliable behavior coding and reinterview data.

Table 3 (located after the references) shows the three models we used to test the predictive utility of the behavior coding data. The dependent variable is the index of inconsistency, a continuous variable that, in theory, ranges from 0 to 100. All three models include the two independent variables for the behavior coding data. These variables are percentages ranging from 0 to 100.[9] The respondent behavior code is the percentage of times respondents provided an adequate or qualified answer to the question. The interviewer behavior code is the percentage of times interviewers read the question exactly as worded or with only slight changes that didn't affect question meaning. In addition to the two behavior coding variables, Model 2 includes three dummy variables representing the sections of the questionnaire. Although the questionnaire contains five sections, two of them--food sufficiency and coping mechanisms and food scarcity--are similar in content and are differentiated in the questionnaire only because the former is used to screen respondents either into or out of the remainder of the questions. Accordingly, these two sections were collapsed for the present analysis. The omitted category is the concern about food sufficiency section. The sections of the questionnaire were included in the model since we knew from both the behavior coding data and the reinterview data that not all of the sections performed equally well. Model 3 includes interactions between the respondent behavior code and the sections of the questionnaire.

Model 1 indicates that the respondent behavior code significantly predicts the index of inconsistency. The sign of the parameter estimate is in the expected direction; that is, as the percentage of respondents who provide adequate or qualified answers increases, the index of inconsistency decreases, indicating lower response variance (higher reliability). Interviewer behavior, however, is not significantly related to the index of inconsistency. These results are similar to those found by Belli and Lepkowski (1995).

The lack of association between interviewer behaviors and question reliability is not surprising. Very few questions were identified as problematic based on interviewer reading errors.[10] Using the 85 percent threshold for determining whether a question was problematic indicates that only 2 of the 37 questions would be considered problematic based on interviewer reading errors. These same two questions plus an additional 12 were determined to be problematic based on respondent codes.

Model 2 includes the dummy variables for the sections of the questionnaire. (The omitted category is the concern about food sufficiency section.) The two behavior coding variables perform similarly in Model 2 as in Model 1. The parameter estimate for the respondent behavior code remains significant and inversely correlated with the dependent variable, and the interviewer behavior codes are not significant. Addition of the three dummy variables contributed significantly to the model R^2. The results indicate that questions in the food expenditures section were associated with higher levels of response variance (more unreliable) and questions in the program participation section were associated with lower levels of response variance (more reliable) than questions in the omitted section. These findings are consistent with the behavior coding data. Using the 85 percent threshold, five of the seven questions from the food expenditures section of the questionnaire that are included in this analysis were identified as problematic based on respondent codes, whereas only one of the five questions in the program participation section of the questionnaire was identified as problematic based on respondent behavior codes.

Model 3 includes interaction terms between the respondent behavior coding data and the section of the questionnaire. The increase in the R^2 value between Model 2 and Model 3 is significant, indicating that the interaction terms contribute significantly to the amount of variation explained in the dependent variable. The interaction terms indicate that the ability of the respondent code to predict the dependent variable is contingent on the section of the questionnaire. The respondent code is significantly associated with the index of inconsistency only in the food expenditures and program participation sections. The respondent code was not significantly associated with the index in the combined food sufficiency/coping mechanisms sections. The questions in this section performed well according to respondent behavior coding data, but produced relatively unreliable data according to the index. And respondent behavior coding data for the concern about food sufficiency section were mixed, whereas the index indicated the questions were uniformly unreliable.

DISCUSSION

Why does behavior coding predict reliability of response in some sections of the questionnaire but not in others? On a purely statistical level, the lack of variation in the independent variable (respondent behavior code) in the combined food sufficiency/coping mechanisms and food scarcity section or the dependent variable in the concern about food sufficiency section is probably sufficient to preclude a significant effect of the behavior coding variable in those sections. The more interesting question, however, has to do with how these sections of the questionnaire differ from the others either in terms of the content of the questions, or in terms of their structure.

One way in which these sections differ from the others is that questions in the food expenditures and program participation sections are of a more clearly factual nature than those in other sections. The food expenditure section includes questions on whether the respondent shopped at various locations (supermarkets and grocery stores, other stores, and restaurants), whether they included all purchases regardless of how they paid for them, how often they shop at supermarkets and grocery stores, and whether the amount they spent last week is the usual amount they spend per week. The program participation questions ask about food stamp recipiency, and participation in other food-related programs such as the school lunch and breakfast program and WIC. The remainder of the questionnaire measures the extent of food insecurity in the household. Questions in the concern about food sufficiency section are intended to measure a more subjective dimension of food insecurity than questions in the food sufficiency/coping mechanisms section. However, one could argue that several of the questions in the latter section are subjective as well.

A second difference is the reference period used in the questions. The food expenditure questions ask about shopping "last week," and the program participation questions ask about the "last 30 days." Questions in the other sections of the questionnaire have either long or nonexistent reference periods. Out of 25 questions, 19 ask about the "past 12 months," 3 ask about the "past 30 days," and 3 mention no reference period. Perhaps the long reference period results in respondents using recall strategies that produce unreliable data. Unfortunately, the data collected in this study do not allow us to investigate these hypotheses further.

CONCLUSIONS

For a long time, researchers have used behavior coding as a guide in questionnaire development, on the assumption that when respondents and interviewers are able to ask and answer questions without difficulty, the quality of the information obtained will be better. This assumption has been based largely on faith rather than empirical evidence. The findings in the present paper provide empirical support for the assumption, but they also appear to qualify it in some important respects. First, interviewer behavior coding has no predictive value for reliability, at least in a study such as this one, where interviewers perform at a uniformly high level. These findings might well differ in studies with greater variability among interviewers. Second, respondent behavior coding data do not appear to predict all types of reliability equally well. Prediction appears to be better for factual questions, and/or for questions with a relatively short recall period. When these conditions are not met, people may be able to answer the questions--and, therefore, behavior coding data may give no indication of difficulty--but the reliability of answers (and, hence, their validity) may nevertheless be low. Clearly, more research is needed into the characteristics of questions for which behavior coding is a valid predictor of test-retest reliability.

In concluding, we would also like to draw attention to some limitations of our data that make us offer these conclusions with a great deal of caution. First, our results are not generalizable. The behavior coding data were not drawn from a random sample of households. They are primarily low income households from the first 75 low income cases interviewed at two of the Census Bureau's centralized telephone facilities. Moreover, the samples for behavior coding and reinterview are different. The reinterview sample is nationally representative, but was oversampled for low income households and suffers from a low response rate (64 percent). Second, because of differences in sample design and sample size, our analysis is at the question level, not the individual level. This analysis would be more precise if we had matched individual level data. Third, the number and type of questions contained in this analysis are very small and the questions are not constructed to deliberately vary either content or structure. Although there were 80 questions in the original survey, we were only able to include 37 questions in our model. Questions were excluded primarily because the characteristic of interest is so rare in the population that the reinterview sample was too small to produce a reliable index of inconsistency. Moreover, we had to exclude continuous variables from the model because the index is calculated differently for categorical and continuous variables and there were too few continuous variables to produce a separate model. Fourth, although approximately 90 percent of the reinterviews were done within seven days of the original interview, the elapsed time between the original interview and the reinterview may account for some of the unreliability measured in the index of inconsistency, and the impact of the elapsed time may not affect all questions equally. It is possible that questions with shorter reference periods, such as those asking about behaviors occurring "last week" in the food expenditures section, were more adversely affected by the elapsed time between interviews than questions with longer reference periods. Respondents may be answering the food expenditure questions about a different week during the reinterview than in the original interview.[11] Thus, the index may not be speaking to reliability in the food expenditure questions and may be correlating with the behavior coding data for the wrong reason. Given these caveats, our results suggest that respondent behavior coding is associated with one measure of reliability; however, its ability to predict reliability in our study was not uniform throughout the questionnaire. Additional research is needed to understand the characteristics of questions for which behavior coding is a valid indicator of reliability and those for which it is not.

[1] Our measure of "185 percent of poverty" in this survey is based on family size and family income. The measure, however, is rather imprecise, because the only measure of family income in the CPS is based on a single question about family income in the previous calendar year and is a categorical variable composed of income ranges.

[2] Race of the household is measured by the unweighted race of the reference person. The reference person is the first person listed on the household roster and is the name of the person or one of the persons who owns or rents the house/apartment.

[3] Contact the authors for a copy of the questionnaire.

[4] We used 5 interviewer codes: 1) exact question reading, 2) slight change in question reading, 3) major change in question reading, 4) verified answer, 5) other. We used 8 respondent codes: 1) adequate answer, 2) qualified answer, 3) inadequate answer, 4) requests clarification, 5) interrupts question reading, 6) don't know answer, 7) refuses to answer, 8) other.

[5] The number of days between the original interview and the reinterview may account for some of the unreliability measured in the index of inconsistency.

[6] Approximately 35 percent of the cases in the original interview were conducted by personal visit and 65 percent were conducted by telephone either from the field representatives' homes or from a centralized telephone facility. Personal visit interviews are primarily month-in-sample one and five cases, that is, those cases that are in sample for the first time or those cases that are returning to the sample after a four-month hiatus. Thus, as much as 35 percent of the sample may be subject to a mode effect and some of the variation in the index may be due to a mode effect. Based on differences in survey data resulting from personal visit vs. telephone mode effects, the consensus at the Census Bureau is that these differences are quite small and would contribute little to the variation in the index.

[7] The index of inconsistency is the simple response variance divided by the total variance. Computationally it is the

proportion who change answers between the original interview and the reinterview divided by (P1*Q2) + (P2*Q1) where P1= the proportion in category from the original interview; where Q1= the proportion not in category from the original interview; where P2= the proportion in category from the reinterview; where Q2= the proportion not in category from the reinterview.

[8] We did, in fact, run a general linear model separately for the numeric data. Because of sample size only the behavior coding variables could be used to predict the index of inconsistency. Neither the respondent nor the interviewer behavior coding variable was significant.

[9] It is possible for the index of inconsistency to be greater than 100 if the number of observed agreements is less than chance. See Perkins, 1971 for details.

[10] Contact the authors for the interviewer and respondent behavior coding data and the index of inconsistency for the 37 questions of interest.

[11] The questionnaire was modified during the reinterview to prompt respondents to report for the week before the original interview.

REFERENCES

Cannell, C., Lawson, S. and Hausser, D. (1975). *A Technique for Evaluating Interviewer Performance*. Ann Arbor, University of Michigan.

Cannell, C., Oksenberg, L., Kalton, G., Bischoping, K., and Fowler, F. (1989). "New Techniques for Pretesting Survey Questions." Report submitted to the National Center for Health Statistics. Ann Arbor, Survey Research Center, University of Michigan.

Hess, J., Singer, E., Ciochetto, S., "Evaluation of the April 1995 Food Security Supplement to the Current Population Survey." Report prepared by the U.S. Bureau of the Census, Center for Survey Methods Research for the U.S. Department of Agriculture Food and Consumer Service, Alexandria, VA, January 26, 1996.

McGuinness, R., "Reinterview Report: Response Variance in the 1995 Food Security Supplement." Report prepared by the U.S. Bureau of the Census, Demographic Statistical Methods Division/QAEB for the U.S. Department of Agriculture Food and Consumer Service, Alexandria, VA, forthcoming.

Perkins, Walter M., "On the Index of Inconsistency." Memo for the Center for Research and Measurement Methods, U.S. Bureau of the Census, 1971.

Presser, S., and Blair, J. "Survey Pretesting: Do Different Methods Produce Different Results?" *Sociological Methodology*, Vol. 2, No. 12, pp. 73-104.

Table 3. General Linear Models for Predicting the Index of Inconsistency (Standard errors in parentheses)

	Model 1	Model 2	Model 3
Variable	Parameter Estimate	Parameter Estimate	Parameter Estimate
Intercept	155.7 (57.1)	76.7 (48.0)	-4.9 (69.0)
Respondent behavior code (RBC)	-0.6* (0.2)	-0.5* (0.2)	0.3 (0.8)
Interviewer behavior code	-0.6 (0.6)	0.2 (0.5)	0.4 (0.4)
Food expenditure (Food)		15.3* (6.8)	268.7** (75.5)
Program participation (Program)		-26.5** (7.7)	201.1* (91.0)
Food sufficiency, coping mechanisms and food scarcity (Coping)		-7.5 (6.5)	34.5 (67.4)
RBC*Food			-3.1** (0.9)
RBC*Program			-2.7* (1.1)
RBC*Coping			-0.5 (0.8)
Model r-square	0.20*	0.61**	0.83**
Degrees of freedom	2	5	8
N	37	37	37

: p<.01 *: p<.05