Journal of Official Statistics
Vol. 8. No. 3, 1992, pp. 293–307
© Statistics Sweden

Question Characteristics and Interviewer Effects

Thomas W. Mangione[1], *Floyd J. Fowler, Jr.*[1], *and Thomas A. Louis*[2]

Abstract: As part of a large study of how interviewers affect data, this paper reports the results of analyses of how question characteristics affect the likelihood that interviewers will influence survey responses.

It was found that ratings of the sensitivity and difficulty of questions were not significantly related to interviewer effects as measured by the intraclass correlation. Whether questions pertained to objective or subjective topics and whether they called for open-ended or fixed responses also were unrelated to interviewer effects. However, it was found that interviewer effects were directly related to the rate at which interviewers had to use follow-up probes in order to obtain an adequate answer. Questions prone to recording errors were also more likely to be affected by interviewers.

The results imply that, during the pretesting phase of a project, questions should be evaluated to identify those that routinely require follow-up probes and redesigned to minimize the need for interviewer initiative to obtain adequate answers.

Key words: Survey questions; interviewer effects; survey methods.

1. Introduction

Nonsampling error is an important, albeit sometimes neglected, concern in the design of sample surveys and the reporting of survey findings. One major type of nonsampling error is interviewer induced error. Although some research has focused on characteristics and behaviors of interviewers which contribute to this error, little research has been conducted which identifies the role item properties have in contributing to interviewer related error.

Kish (1962) used the intraclass correlation, p, to measure interviewer effects. A perfectly standardized set of interviewers will contribute nothing to item variance; instead the variation will come from true score variation and random error due to item properties, respondent characteristics, or situational factors. None of the error would be correlated with the interviewer. However, to the extent that interviewers are not standardized, and hence influence the data, their effect will be observed as error correlated with the interviewer and can be assessed by p. We calculate p by comparing the observed total error to the mean square error term (or error due to simple random sampling).

The significance of p for total survey error calculations is its effects on the calculation

[1] Center for Survey Research, University of Massachusetts, 100 Arlington Street, Boston, MA 02116, U.S.A.
[2] Division of Biometry, School of Public Health, University of Minnesota.
Correspondence to: Thomas W. Mangione, Center for Survey Research, 100 Arlington Street, Boston, MA 02116, U.S.A.

Acknowledgement: The research presented here was supported by the National Center for Health Services Research, Grant H504189.

of standard errors. One measure of total survey error is the design effect

$$\text{DEFT} = \sqrt{1 + (n - 1)p} \qquad (1.1)$$

where n is the average size of the interviewers' assignments. The square root of the design effect (sometimes called DEFT) represents the proportionate increase in the size of the standard error due to the effect of interviewers on the data they collect over and above the effects of the sample design.

Historically the interviewer has been thought to be a force both to create and to reduce nonsampling errors. On the one hand, interviewers can probe unclear or incomplete answers, clarifying points of confusion or definition, and ensure that respondents meet question objectives. However, early research by Hyman, Cobb, Feldman, and Stember (1954) demonstrated that interviewers increased error variance through nonstandardized interviewing procedures. Subsequent research by others has shown that interviewers are associated with error. Collins (1980) showed that when interviewers were given discretion about reading a neutral response option, very high interviewer-related error occurred. Rustemeyer (1977) found interviewers made recording errors that affected the resulting data in 10% to 14% of the answers they recorded. Weiss (1968) found that interviewers who rated their "rapport" with respondents as high obtained less accurate data when the answers were compared to records. In a series of studies Cannell and his colleagues have found that they can program interviewer behavior and communications in ways that will influence the quality of data reported (Cannell, Fowler, and Marquis 1968; Cannell, Marquis, and Laurent 1977; Cannell, Groves, Magilavy, Mathiowetz, and Miller 1987). Elsewhere, Fowler and Mangione (1990) showed that interviewer effects were associated with the amount of training they received and the types of supervision used with the interviewers.

The number of studies, however, that try to explain why interviewer effects exist is relatively limited. To understand the source of interviewer effects will require investigating the underlying factors.

Previous research shows that there is a wide variation across items as to the size of interviewer effects. Groves (1989) presented data from ten personal interview studies which showed about half of the items with a p of .03 or greater and nine telephone interview studies with about half of the items with a p of .009 or greater. Tucker (1983) used 11 national telephone polls and showed that about half the items had a p of .004 or greater. As a frame of reference a p of .01 with an average interviewer load of 25 interviews will result in an 11% increase in estimates of standard errors.

In addition to noting the prevalence of survey items with sizable interviewer effects, it is intriguing to also note that not all items are affected. The question can be posed: Are there some types of items that are distinctly susceptible to interviewer effects? Stated somewhat differently: Are there some types of items that create the climate for interviewer effects to occur?

Although interviewer effects have been identified in the methodological literature for a long time, there are very few, if any, well supported generalizations about the conditions under which significant interviewer effects occur or how to minimize them. (See Groves 1989; and Stokes and Yeh 1988). There is limited support in the literature for three aspects of item content to influence interviewer effects – item difficulty, sensitivity, and whether the item asks about factual matters. Research by Cannell et al. (1977) has shown that events that are of minor importance or more distant in time (and hence harder to remember) are reported with less accuracy than average; Hansen, Hurwitz, and Bershad (1961) found difficult items such as income and occupation more susceptible to interviewer effects.

Research reported by Bradburn, Sudman and Associates (1979) and Locander, Sudman, and Bradburn (1976) and others have shown that underreporting is a common problem with sensitive topics and Cannell et al. (1977) report that hospitalizations for "threatening" conditions are underreported. Marquis, Marquis, and Polich (1986) linked item reliability to question sensitivity. Fellegi (1964) found emotionally charged items more susceptible to interviewer effects. Many of these studies focused on response error, however, and response error per se does not mean that the error is interviewer related. Moreover, the conclusions about the relationships between item content and p have mainly been based on post hoc analyses involving only a few items.

Most questions can be coded as dealing with either facts or opinions. The answers to questions about people's feelings or ideas cannot be directly verified. In contrast, questions about behaviors or events could, in theory, be measured independently from the respondent. Therefore, factual questions have an objective anchor that may make their reporting less susceptible to interviewer influence than opinion items. Kish (1962) did not find major differences between factual and opinion items; nor did Groves (1989) find major differences in analyses across nine studies with 297 questions. Others (Hansen et al. 1961; Fellegi 1964; Feather 1973; O'Muircheartaigh 1976; Collins and Butcher 1982), however, have reported that factual items are less susceptible to interviewer effects.

In addition to these three dimensions of item content or wording, one aspect of item form could be considered: whether the question called for an open-ended response or asked respondents to choose one of a list of answers provided (closed response). Open response questions demand more active involvement by the interviewer to obtain complete answers, hence, there is more opportunity for influence. Collins (1980), Gray (1956), and O'Muircheartaigh (1976) showed interviewers were more likely to affect the answers to open response questions. Groves and Magilavy (1986) reported that interviewers were particularly likely to affect whether or not a codable answer was obtained, and the number of points made in response to such questions.

An experimental study of the value of interviewer training and supervision (Fowler and Mangione 1986) also provided an opportunity to study the characteristics of questions that make them susceptible to interviewer effects. Results of analyses of interviewer characteristics and behaviors are published elsewhere (e.g., Fowler and Mangione 1986, 1990). This paper focuses on three main issues. (1) It describes the distribution of p in our study and how it compares to prior research. (2) It presents our *a priori* testing of the effect of item characteristics on the size of p. We focused on three dimensions of question content that we felt would potentially produce susceptibility to interviewer effects: the level of difficulty of the question; the potential sensitivity of answers; and whether it was a factual or an opinion question. We also tested the effect of item form (open versus closed questions). (3) It presents our *post hoc* analyses of item characteristics and interviewer behaviors that are associated with high values of p.

2. Methods

2.1. *Overview of design*

Fifty-seven newly recruited interviewers were randomly assigned to one of four training programs, ranging from one to ten days and to one of three different programs of supervision. They then were each given an assignment of 40 addresses in suburban Boston. Each interviewer's assignment was a probability subsample of the total sample, so that differences in response patterns of a given interviewers' respondents, beyond normal sampling variability, could be attributed to the interviewer.

Interviewers were to interview an objectively designated adult in each sampled household. The in-person interviews averaged about 30 minutes. The interview consisted of a carefully constructed set of health services questions, the characteristics of which are described below. No crossover of sample to other interviewers was allowed. The average response rate for all interviewers was 67%.

2.2. *Item classification*

The first step in the process of questionnaire design was to identify various dimensions of items that were either of theoretical or practical significance. We chose to focus on the difficulty of an item, its sensitivity, whether it was opinion or factual, and whether it was open or closed ended.

We defined a difficult item as one which:

> required the respondent to recall things that may be hard to remember (i.e., minor events, counting things over a period of time unless it was a very short time) *or* dealt with an issue that was complicated *or* the respondent was unlikely to have thought much about before.

Sensitive items were defined as ones for which:

> there was a fairly pervasive norm such that giving a particular answer would make the respondent look better or would be more socially acceptable.

Factual items were defined as ones for which:

> there was the potential ability to corroborate the answer from an independent source.

Open items were defined as ones for which:

> the response alternatives were not read to the respondent but instead they were allowed to answer in their own words. Items that asked for numerical quantities and for which respondents were not given categories were also counted as open questions.

The judgments of whether items were sensitive or difficult presented a great deal of problems for staff since they involved subjective judgments. Individual ratings were made by four staff members and then discrepancies were resolved by group discussion and consensus. The strength of this approach was its *a priori* coding of item type which prevented *post hoc* reasoning from unduly influencing the creation of our independent variables.

2.3. *Questionnaire construction*

The content of the interview schedule was typical of health services research: use of services, health status, health behaviors, mental health, health policies, and background questions. However, item construction proceeded in the reverse of the usual process. Form and content dimensions were mapped first, and then the subject matter was created. The process actually became interactive, since the questions eventually had to fit together into a questionnaire with sections dealing with various health areas.

There were 16 cells formed by the combi-

Table 1. Number of items included in questionnaire with various characteristics out of a total of 130 items

Characteristic	Number of items
Difficult/Easy	65/65
Sensitive/Not sensitive	84/46
Opinion/Factual	45/85
Open/Closed	50/80

nation of the four question characteristics in which we were interested. Examples were gathered of the most common questions used in health services research surveys sorted by their characteristics. Further questions were added or deleted to create a cohesive questionnaire that had a logical flow; other questions were inserted to round out the substantive value of sections of the questionnaire.

When the questionnaire was completed, all questions were coded in their final classification by the project staff (again independently and then as a group to resolve disparities) and allocated or (reallocated) to appropriate cells. Table 1 shows the characteristics of the 130 items as finally coded. Although the combination of characteristics was not perfectly balanced, there was a good representation of questions in all the main categories.

Table 2 shows examples of questions that were used in the questionnaire. One question is shown for each of the 16 cells, which represent all combinations of the four question characteristics we assessed.

2.4. *Calculation of p for each variable*

A critical step for this analysis was the calculation of an intraclass correlation (p) for each survey item. The first step was to examine the distribution of answers for each ordinal and interval scale item for a reasonable approximation to a normal distribution. For items which deviated markedly (e.g., a highly skewed distribution) a transformation was made to more closely approximate a normal distribution (e.g., taking the square root).

For the nominal scale variables with more than two answer categories, the procedure was more involved. For each category within a variable that represented at least 5% of the answers, a dummy variable was created, and p was calculated for each such category. For each question, an average p was calculated by averaging the p for each answer category. This average p was used in the analyses below. This method was considered more desirable than four alternatives: (1) arbitrarily picking one answer to represent the questions, or (2) picking the most frequently mentioned answer, or (3) using all the answers (which would have artificially increased the number of estimates in the analysis), or (4) using the highest value of p for a given question.

These procedures created 130 variables for which a p could be calculated. One-way analyses of variance were run, using the SAS General Linear Models subroutine, with "interviewer" as the independent variable (56 degrees of freedom), for each of the 130 variables.

These ANOVAs produced statistics which enabled a calculation of p for each variable using the following formula

$$p = \frac{\frac{V_a - V_b}{m}}{\frac{V_a - V_b}{m} + V_b} \qquad (2.1)$$

where V_a is the between mean square in a one-way analysis of variance with interviewer as the factor; V_b is the within mean square in the analysis of variance; and m is the average total number of interviews conducted by each of the interviewers.

Some calculated values of p were negative. For the analyses presented in this paper, these values were recoded to a very small positive value (.001) consistent with

Table 2. Examples of items by dimensions of form and content

Difficult	
Not Sensitive	Sensitive
Opinion–Open: From what source would you say you get the most information about health and what you should do to keep healthy?	*Opinion–Open:* Under which circumstances, if any, do you think a woman should be legally permitted to have an abortion?
Opinion–Closed: In general, when people have personal or family problems, do you think it is better for them to get professional counseling right away or better for them to try to work their problems out on their own?	*Opinion–Closed:* If (you/your wife) were pregnant and the doctors told you that it was almost certain the baby would be born with a serious deformity, how likely is it that you would decide to end the pregnancy with an abortion – very likely, fairly likely, or is there no chance at all?
Factual–Open: How many days in the past year did you stay in bed all or most of the day because of any illness or injury?	*Factual–Open:* How many different days have you had any beer, wine or liquor to drink in the last 30 days – that is since (DATE) a month ago?
Factual–Closed: In the past year, how much would you say you spent out of your own pocket on your *own* medical care. Count hospital and doctor bills, bills for prescription medicine supplies and tests; don't count any dental costs. Would you say your costs in the past year were nothing, less than $50, $50 to $100, $100 to $250, or over $250?	*Factual–Closed:* Are you able to run or jog half a mile without stopping?

Table 2 (cont.). Examples of items by dimensions of form and content

Easy	
Not Sensitive	Sensitive
Opinion–Open: Coffee, tea and cola soft drinks are popular beverages that contain caffeine. What is your favorite beverage that has caffeine in it?	*Opinion–Open:* (Was there anything that happened yesterday with your family, friends, work or whatever, that made you feel particularly bad – that worried or depressed you or made you upset) IF YES: What happened?
Opinion–Closed: How important do you think it is for a person your age to have a general physical check-up every year – very important, somewhat important or not at all important?	*Opinion–Closed:* How would you rate the way you take care of your health – excellent, fair or poor?
Factual–Open: What is your height?	*Factual–Open:* How many cigarettes did you smoke yesterday?
Factual–Closed: How long ago was the last time you were actually seen by a doctor about your health – within the last month, 1 to 6 months ago, 6 months to a year ago, or more than a year ago?	*Factual–Closed:* Have you ever smoked marijuana?

approaches used by other researchers (e.g., Bailar, Bailey and Stevens 1977). The value of p for each variable was then transformed by the following equation to produce better distribution characteristics for analysis

$$\text{transformed } p = \log(p/(1-p)). \quad (2.2)$$

To check that the reduction in variance produced by recoding negative values did not distort results, the analyses presented here were also replicated using untransformed values of p; the results were virtually identical.

An item file was created which had 130 records, each of which contained the value of p transformed, and the coding of item characteristics. Item characteristics, initially coded "0" or "1", were transformed by subtracting the mean of each variable from each category. This maintained a distance of one unit between the categories but resulted in a mean score of zero for each characteristic, and thereby created "centered" independent variables.

2.5. Interviewer-respondent interaction coding

A part of the basic experimental study involved tape recording interviews. A third of the study interviewers tape recorded most of their interviews, and all interviewers taped at least their first interview (the respondent for this first interview was not one of the sample elements and hence could be considered to be a practice interview. From the interviewer's perspective, however, all interviewing procedures were identical to those for the rest of the sample).

After the initial analyses described in this paper were completed, we became interested in the particular interviewer behaviors that could be associated with items that showed interviewer effects. To study this topic, we used 100 tape recorded interviews that were available at the time (at least one from each of the 57 interviewers and no more than five from any one interviewer).

Building on strategies reported by Cannell et al. (1968) and Guenzel, Berkmans, and Cannell (1983), a detailed coding was made of interviewer behaviors exhibited while trying to obtain answers to 65 questions selected to represent the range of p's calculated as described above. For each question within each taped interview, the occurrence of the following behaviors was noted:

1. Interviewer laughs.
2. Interviewer does not read question exactly as worded.
3. Interviewer uses correct nondirective probe.
4. Interviewer uses directive probe.
5. Interviewer fails to probe an inadequate answer.
6. Interviewer provides evaluative (and hence inappropriate) feedback on answer given.
7. Interviewer initiates inappropriate interpersonal behavior.
8. Interviewer fails to record answer to open-ended question accurately and verbatim.

It is important to understand that our goal was to find out how the questions affected the way interviewers behaved. Cannell, Oksenberg, and Kalton (1991) have recently shown, what we suspected, that questions have consistent effects on behavior in interviews. From this perspective, the behavior coding was used to identify properties of the questions that served as the behavioral stimuli.

The coders were carefully trained and check coded until their level of agreement exceeded 90%. The relationships between these coded behaviors and the values of p are examined in Section 3.2.

3. Results

3.1. Question characteristics and p

The cumulative distribution of the calculated p from the 130 items in this health survey showed reasonable variation in the size of p. About 35% of the items showed no interviewer effect; another 35% of the items showed only modest effects with p less than .01. About 20% of the items showed p ranging from .01 to .023. Finally, 10% of the items showed p in excess of .023 with a maximum value for these items of .049. The value of p needed for statistical significance depended on the number of respondents answering the question, but generally p's in excess of .015 were significant at the .05 confidence level.

This distribution of p across items compares to similar distributions found in earlier studies by Kish (1962) and Hanson and Marks (1958), but showed somewhat fewer extreme interviewer effects than found in studies by Freeman and Butler (1976), Collins (1980), and Groves and Kahn (1979). In part, this may be due to our procedure of calculating an average p for nominal scale variables.

Table 3 lists the question wording of the 13 items in our study with the highest p's. Items with p's in this range with an average interviewer load of 25 interviews would inflate standard errors from 25% to 48%.

In order to determine the effect of item characteristics on the size of the interviewer effects, regression analyses were run on the

Table 3. Wording of those questions with the highest interviewer effects (p)

	p	Item
1.	.049	What is the main reason you would probably go to that hospital rather than some other hospital?
2.	.046	What is the main reason you would probably go to that hospital for serious surgery?
3.	.038	How long ago was the last time you were actually seen by a doctor for your health – within *the last month*, *1 to 6 months ago*, *6 months to a year ago*, or *more than a year ago*? (Probe: about how many years ago was that?)
4.	.037	Are you not working because you are *unemployed*, *on layoff from a job*, *retired*, *a student*, *keeping house*, or what?
5.	.035	In the past 12 months did you have eczema or psoriasis?
6.	.034	How many days in the last month would you say you had (# drinks *R* reported usually had) drinks?
7.	.032	On those days last month when you drank, how many drinks of beer, wine, or liquor did you usually have? Count a can of beer, a glass of wine, or $1\frac{1}{2}$ ounces of liquor as a drink.
8.	.029	What kind of place is that (where you usually go for health care) – a *clinic*, *health center*, a *hospital*, a *doctor's office*, or *some other place*?
9.	.026	Why did you go to the doctor the last time you went?
10.	.025	How long do you usually (main form of exercise) when you do it?
11.	.024	(Males only) In the past 12 months did you have prostate trouble?
12.	.024	Another health-related issue is the conditions under which abortions should be performed. Is that a topic you have thought about – *a lot*, *some*, or *only a little*?
13.	.023	In the past 12 months did you have (a) migraine?

transformed value of p. A stepwise, forward inclusion regression procedure was used. The main effects were represented by the centered variables. Four item characteristics were included (difficult, sensitive, opinion, and open) as main effects. Two-way and three-way interaction terms were created by multiplying the centered main effect variables. The main effects were forced in first, followed by the two-way interactions and three-way interaction terms. Within each group, variables were allowed to enter as determined by the strength of their relationship to p.

The result of the regression procedure is shown in Table 4. The use of the centered independent variables allows the constant term to be interpreted as the average value of p for this sample and the effect of each variable to be directly interpreted as the deviation from the average level of p for the property or properties that it represented.

Overall, the model had a multiple R of .35, which corresponded to 12% of the variance in p accounted for by item characteristics. The probability of this model being different from no effect was only .13.

The only main effects that even approached statistical significance were the level of difficulty of an item or whether the item was open. The more difficult an item was ($p = .11$), or whether the item was open ($p = .10$), then the more susceptible to interviewer effects it was. Neither of the other main effects showed a significant association with p.

When interactions between the item properties were looked at, the open-sensitive interaction term was the only one that reached the .05 level of significance. Contrary to expectations, however, open-sensitive items were less susceptible to interviewer effects and open-non-sensitive were more susceptible to interviewer effects. This significant interaction term means that interviewers affected answers more for open

Table 4. Stepwise regression of item characteristics on transformed p's

	Only main effects entered		Final step	
Item characteristics	B	Sig. of F	B	Sig. of F
Difficult	+.199	.11	+.408	.11
Sensitive	−.115	.27	+.102	.52
Opinion	−.013	.92	−.177	.51
Open	+.028	.79	+.334	.10
Combinations				
Open–Sensitive			−.474	.04
Opinion–Difficult			+.641	.11
Opinion–Sensitive			+.084	.86
Difficult–Sensitive			+.256	.39
Open–Opinion			+.356	.18
Difficult–Open			−.284	.27
Opinion–Difficult–Sensitive			−.503	.37
Constant	−2.355		−2.435	
R	.21		.35	
R^2 (R^2 Adjusted)	.04 (.01)		.12 (.04)	
Sig of Model	.24 (4,125)		.13 (11,118)	

non-sensitive items than open-sensitive items or for any closed item. It was as if interviewers were less diligent when dealing with non sensitive, open questions but were on their best behavior when dealing with sensitive, open questions. Corroborating interviewer behavioral data are presented in the next section.

3.2. Interviewer behaviors and p

Although there were some limitations to our analyses which we will discuss below, the fact remains that the results from the regression analysis show low correlations with our coding of item characteristics. This led us to look directly at interviewer behaviors to see if we could determine how an item caused an interviewer to influence respondents and to determine whether there were properties of items different from those tested which were associated with interviewer effects. A review of taped interviews from each interviewer for a sample of items forms the basis for our next set of findings.

Table 5 shows the correlations between various interviewer behaviors and p. The measures of behavior focus on those aspects of the interviewing process which were observable from the taped interviews and which could be a source of influence on the respondents' answers. We included measures of probing quality, correct presentation of the question, measures of feedback and interpersonal interactions, and accuracy of recording.

The findings presented reflected our hypotheses that questions with high p values were likely to induce directive probes or failures to probe. However, the fact that all the probing-related codes were positively associated with levels of p means that the real issue is the likelihood that respondents will give an initial inadequate answer to high p items and these answers require probing. As Cannell et al. (1991) found, some questions consistently produce answers that do not meet question objectives and require probing. Such questions will produce higher than average rates of all probing-related behaviors. Since probing requires interviewer discretion, the result, as Table 5 shows, is higher than average values of p.

In addition, inaccurate verbatim recording on open-ended questions also was associated with high interviewer effects. Overall, not all open questions were susceptible to interviewer effects, as we have seen previously. However, those where interviewers had trouble recording fully and accurately had higher values of p. Also, it was clear that the poor recording was not merely a difference between recording exact words versus summaries, but the errors resulted in substantively different answers being coded given the way the p's were calculated for open questions.

None of the interpersonal behaviors coded were correlated with the size of p. This was no doubt in part due to the fact that such behaviors occurred in less than 1% of the interactions on particular questions.

The rate of incorrect reading of the question was not correlated with the size of p. About 17% of the items were misread, so infrequency of occurrence was not the explanation for the low association. Our coding did not differentiate between significant and insignificant misreadings; it is plausible that some changes in question wording would affect the answers more than other changes would.

Table 6 shows the relationship between our coding of item characteristics and the observed interviewer behaviors. From this table it was clear that difficult items, opinion items, and open items are more likely to cause interviewers to exhibit incorrect interviewer behavior on almost all of the dimensions measured. In particular, these items

Table 5. Significant correlations between the incidence of specific interviewer behaviors and p

Interviewer behavior	Correlation with p
Laughing	–
Incorrect reading of question	–
Correct probe	.23
Directive probe	.20
Failed to probe	.49
Inappropriate feedback on answer	–
Inappropriate interpersonal behavior	–
Incomplete or inaccurate verbatim recording on open Q's	.39

Note: Cells with dashes showed nonsignificant correlations, $p < .05$. Correlations are based on coding of behaviors while asking 65 items for 100 different interviews. Degrees of freedom = 64.

require more probing and hence probing problems were likely to be exhibited on these items.

Items about sensitive topics were not associated with undesirable interviewer behaviors. If anything, there was some evidence that interviewers were particularly careful in how they dealt with these items. Sensitive items were more likely than average to be probed when required, and interviewers recorded verbatim answers more completely than on average.

4. Discussion

To the extent that there has been discussion in the literature about question characteristics that produce interviewer effects, it has tended to focus on the content of the item. Hence, the most pervasive hypotheses are that sensitive items, difficult items, and attitudinal items might be most subject to interviewer effects. The fact that we could find little discernible relationship between these characteristics and the level of interviewer effects, with the possible exception of

Table 6. Significant correlations between question characteristics and the incidence of specific interviewer behaviors

Interviewer behavior	Difficult	Sensitive	Opinion	Open
Laughing	.34	–	.31	.35
Incorrect reading of question	.52	–	.45	–
Correct probe	.34	–	.24	.63
Directive probe	.59	–	.57	.48
Failed to probe	.28	–.22	.38	.56
Inappropriate feedback on answer	.29	–	.24	.23
Incomplete or inaccurate verbatim recording on open Q's	.28	–.26	.38	.59

Note: Cells with dashes showed nonsignificant correlations, $p < .05$. Correlations are based on coding of behaviors while asking 65 items for 100 different interviews. Degrees of freedom = 64.

difficult items, is good news for researchers. It means that interviewer effects are not a necessary adjunct to questions on any particular topic.

These analyses benefited from the procedures that applied *a priori* coding of question characteristics. Even so, the most significant limitation of our study is, we readily acknowledge, the subjective nature of the judgments in coding of difficult or sensitive items. We also found these judgments difficult to make. Although we sought reliability through consensus, the unreliability of our coding cannot be ruled out as a factor which diminished the strength of our associations.

A key issue in the coding of sensitive or difficult items is that it is actually the answer, not the question *per se*, that determines whether or not a question is sensitive or difficult. For instance, questions about drug use are not problems for people who have never used drugs. We will say that our effort at coding question properties prior to the analyses, though possibly imperfect, is an improvement in method over most previous analyses. We would welcome other attempts to study these issues. In addition, we should note that our results are reasonably robust; when we recoded a few sensitive items because we were unsure of our coding, the results were the same.

Our measure of interviewer effects, *p*, also has its limitations. It only measures inconsistency across interviewers in ways that affect data. It does not capture biasing effects. This may be particularly important when considering sensitive items. Our results showed relatively little inconsistency among interviewers in administering sensitive items; we were not, however, testing whether these same interviewers were introducing any bias to the answers received.

The most important finding of our analyses, however, is that questions that require interviewers to probe are those that are most subject to interviewer effects. Our findings are consistent with Groves and Magilavy (1986) where they found the number of responses to open questions to be correlated with interviewers. To our knowledge, our findings are the first empirical documentation of the link between ease of interviewer administration and the quality of data that results. Although those responsible for managing interviewers have long argued for clear questions that can be read exactly as worded and that prepare respondents to give answers, survey research is rife with questions that are difficult to administer and to answer. The data from these analyses suggest that such questions, questions that interviewers cannot simply ask once and obtain an adequate answer, are not only difficult for interviewers but also increase the error in survey estimates.

Despite the fact that there is more research to be done, we think these analyses have at least four implications for researchers.

1. Interviewer assignment size should be kept reasonably small. Ultimately, interviewer effects are dependent on the size of the assignment as well as *p*, and the effect can be lowered by keeping interviewer assignment sizes low. This is particularly important for telephone studies where it is not unusual to see interviewers completing in excess of 100 interviews.

2. Interviewers should be well trained in probing and in verbatim recording.

3. Quality control while a study is in progress should include monitoring of the probing and recording behavior of interviewers.

4. The most important implication of these findings is that one way to reduce interviewer effects is to design questions that minimize the need for interviewers to probe in order to produce a usable answer.

We realize few investigators knowingly produce items that cause problems for respondents or interviewers. Nevertheless, designing questions to minimize the need for probing is not an accepted standard in question design. It is clear from our data that any question that routinely requires interviewer probing is one which interviewers are likely to handle in an inconsistent way. A question that must be probed, that respondents do not answer readily after it is read once, simply presents an opportunity for interviewers to be inconsistent across respondents and between interviewers.

From this analysis, it follows that one key to better surveys is better pretesting. Instead of the relatively unsystematic pretests that are common, relying on relatively nonsystematic and unstructured feedback from interviewers, researchers should tape record pretest interviews. These tape recordings can then be coded to ascertain the rate at which each question had to be reread or required probing. Based on pretest data, questions which stand out in the frequency of interviewer probing required are candidates for revision.

Procedures for better pretests are beginning to appear in the research literature (Cannell et al. 1991; Fowler 1989). By doing a better job of identifying questions that are not clear and adequate, researchers would make the interviewer's job easier, make the question and answer process go more smoothly for respondents, and in addition, would improve the precision of survey-based estimates.

5. References

Bailar, B.A., Bailey, L., and Stevens, J. (1977). Measures of Interviewer Bias and Variance. Journal of Marketing Research, 14, 337–343.

Bradburn, N.M., Sudman, S., and Associates. (1979). Improving Interview Method and Questionnaire Design. San Francisco: Jossey-Bass.

Cannell, C.F., Fowler, F.J., and Marquis, K.H. (1968). The Influence of Interviewer and Respondent Psychological and Behavioral Variables on the Reporting in Household Interviews. Vital and Health Statistics, Series 2, No. 26. Washington: U.S. Government Printing Office.

Cannell, C.F., Groves, R.M., Magilavy, L., Mathiowetz, N., and Miller, P.V. (1987). An Experimental Comparison of Telephone and Personal Health Surveys. Vital and Health Statistics, Series 2, No. 106. Washington: U.S. Government Printing Office.

Cannell, C.F., Marquis, K.H., and Laurent, A. (1977). A Summary of Studies. Vital and Health Statistics, Series 2, No. 69. Washington: U.S. Government Printing Office.

Cannell, C.F., Oksenberg, L., and Kalton, G. (1991). New Strategies for Pretesting Survey Questions. Journal of Official Statistics, 7, 349–365.

Collins, M. (1980). Interviewer Variability: A Review of the Problem. Methodological Working Paper No. 19. London: Social and Community Planning Research.

Collins, M. and Butcher, B. (1982). Interviewer and Clustering Effects in an Attitude Survey. Journal of the Market Research Society, 25, 39–58.

Feather, J. (1973). A Study of Interviewer Variance. Saskatoon, Canada: Department of Social and Preventive Medicine, University of Saskatchewan.

Fellegi, I.P. (1964). Response Variance and Its Estimation. Journal of the American Statistical Association, 59, 1016–1041.

Fowler, F.J., Jr. (1989). Conference Proceedings, Health Survey Research Methods, NCHSR and Health Care Technology Assessment, DHH Publication #89-3447.

Fowler, F.J., Jr. and Mangione, T.W. (1986). Reducing Interviewer Effects of

Health Survey Data. Washington, D.C.: National Center for Health Services Research.

Fowler, F.J., Jr. and Mangione, T.W. (1990). Standardized Survey Interviewing. Beverly Hills: Sage Publications.

Freeman, J. and Butler, E.W. (1976). Some Sources of Interviewer Variance in Surveys. Public Opinion Quarterly, 40, 79–91.

Gray, P.G. (1956). Examples of Interviewer Variability Taken From Two Sample Surveys. Applied Statistics, 5, 73–85.

Groves, R.M. (1989). Survey Errors and Survey Costs. New York: John Wiley.

Groves, R.M. and Kahn, R.L. (1979). Surveys by Telephone: A National Comparison with Personal Interviews. New York: Academic Press.

Groves, R.M. and Magilavy, L.J. (1986). Measuring and Explaining Interviewer Effects in Centralized Telephone Surveys. Public Opinion Quarterly, 50, 251–266.

Guenzel, P.J., Berkmans, T.J., and Cannell, C.F. (1983). General Interviewing Techniques. Ann Arbor, MI: Institute for Social Research.

Hansen, M.H., Hurwitz, W.N., and Bershad, M.A. (1961). Measurement Errors in Censuses and Surveys. Bulletin of the International Statistical Institute, 38, 359–374.

Hanson, R.H. and Marks, E.S. (1958). Influence of the Interviewer on the Accuracy of Survey Results. Journal of the American Statistical Association, 53, 653–655.

Hyman, H., Cobb, J., Feldman, J., and Stember, C. (1954). Interviewing in Social Research. Chicago: University of Chicago Press.

Kish, L. (1962). Studies of Interviewer Variance for Attitudinal Variables. Journal of the American Statistical Association, 57, 92–115.

Locander, W., Sudman, S., and Bradburn, N. (1976). An Investigation of Interview Method, Threat and Response Distortion. Journal of the American Statistical Association, 71, 269–275.

Marquis, K.H., Marquis, S.M., and Polich, J.M. (1986). Response Bias and Reliability in Sensitive Topic Surveys. Journal of the American Statistical Association, 81, 381–389.

O'Muircheartaigh, C.A. (1976). Response Errors in an Attitudinal Sample Survey. Quality and Quantity, 10, 97–115.

Rustemeyer, A. (1977). Measuring Interviewer Performance in Mock Interviews. Proceedings of the Social Statistics Section, American Statistical Association, 341–346.

Stokes, S.L. and Yeh, M., (1988). Searching for Causes of Interviewer Effects in Telephone Surveys. In R. Groves, P. Biemer, L. Lyberg, J. Massey, W. Nicholls, II, and J. Waksberg (eds.), Telephone Survey Methodology, New York: John Wiley, 357–373.

Tucker, C. (1983). Interviewer Effects in Telephone Surveys. Public Opinion Quarterly, 47, 84–95.

Weiss, C.H. (1968). Validity of Welfare Mothers' Interview Responses. Public Opinion Quarterly, 32, 622–633.

Received January 1991
Revised January 1992