AMERICAN JOURNAL OF EPIDEMIOLOGY
Vol. 101, No. 1

Printed in U.S.A.

INTERVIEWER EFFECT ON RESPONSES TO A QUESTIONNAIRE RELATING TO MOOD[1]

IVAN C. CHOI AND GEORGE W. COMSTOCK

Choi, I.C. and G.W. Comstock (Training Center for Public Health Research, Box 2067, Hagerstown, MD 21740). Interviewer effect on responses to a questionnaire relating to mood. *Am J Epidemiol* 101:84–92, 1975.—A community mental health assessment questionnaire relating largely to depressed mood was administered during 1972 to 1,212 respondents in Washington County, MD, by six interviewers. Analysis of 15 psycho-social tests showed that responses obtained by one interviewer differed significantly from responses obtained by the other five. Suggestions for minimizing interviewer effects include 1) selection of interviewers with similar characteristics and backgrounds; 2) adequate training and periodic field assessment of interviewer performance; 3) simplification of questions and reduction in the number of possible responses per question; and 4) allocation of various types of subjects to interviewers as uniformly as possible.

attitude to health; attitude of health personnel, interviewers' effect on responses; health surveys; sampling studies

In spite of the fact that household interviews have long been used in epidemiologic studies, surprisingly little information about some aspects of interview methodology is available in the epidemiologic literature. This is particularly true of the relation of response variation to interviewer characteristics, a subject that has received much more attention in the social science field (1).

Observer variation has been most thoroughly documented by epidemiologists in the area of laboratory determinations. Studies on the reproducibility of red blood cell counts were among the earliest to call attention to observer variation (2). Most thoroughly studied have been differences in the recognition and diagnosis of tuberculosis and cardiovascular disease by interpreters of chest roentgenograms (3–6). Observer variation has also been documented in studies of blood pressure (7–9), serum glucose (10), and clubbing of the fingers (11).

Schilling et al. (12) reported diagnostic disagreements between observers of byssinosis among cotton workers in Lancashire, England. The diagnosis was based on the history of exposure and symptoms, clinical signs of respiratory disease, measurement of chest expansion, and blood

Received for publication May 6, 1974, and in final form August 26, 1974.

[1] From the Community Mental Health Epidemiology Program, Center for Epidemiologic Studies, National Institute of Mental Health; and the Training Center for Public Health Research and the Department of Epidemiology, School of Hygiene and Public Health, Johns Hopkins University, Baltimore, MD 21205.

Reprint requests to Training Center for Public Health Research, Box 2067, Hagerstown, MD 21740.

This study was supported in part by contract HSM 42-71-32 from the National Institute of Mental Health; Residency Training Grant in General Preventive Medicine No. 5-A08-AH 00059, National Institutes of Health; and Research Career Award HL 21,670, National Heart and Lung Institute.

The authors are grateful to Dr. James A. Tonascia, Department of Biostatistics, School of Hygiene and Public Health, The Johns Hopkins University, for assistance and advice with adjustments of the data by multiple regression. They are also indebted to Mr. Knud J. Helsing for many helpful suggestions and to the rest of the field staff of the Community Mental Health Assessment Program in Washington County for their persistently conscientious efforts.

pressure readings. Significant observer variation was noted, particularly in the recording of physical signs.

With respect to interviewer effect on responses to questions, most of the work related to epidemiology has been in the field of chronic respiratory disease. Cochrane et al. (13) first called attention to the important degree of variation between observers in obtaining a history of respiratory symptoms among miners when questions and probes were not standardized. Similar findings were reported by Jonathan et al. (14) from their studies of pneumonoconiosis. Holland and his co-workers (15) compared respiratory histories obtained by two different questionnaires, one developed by the Medical Research Council and the other by the National Coal Board. Both gave reproducible results and were equally discriminating when tested against sputum volume and ventilatory function. However, the two questionnaires yielded different estimates of the prevalence of certain symptoms. Part of the difference was thought to be caused by the wording of the questions; part resulted from the order in which the questions were given. Both findings emphasized the need for standardizing the content and administration of questionnaires used in comparative studies.

A recent report confirmed that interobserver variation could be minimized by the use of the standardized Medical Research Council-European Community for Coal and Steel (MRC-ECCS) respiratory questionnaire (16). However, some inconsistencies were noted in the prevalence of grade 2 breathlessness (such as occurs when "hurrying on the level or walking up a slight hill") which were thought to be related somehow to the way in which the question was stated (17). A useful summarization of the possible sources of disagreement between observers, illustrated by the diagnosis of chronic bronchitis by history, has been given by Fletcher in terms of definition, reproducibility, validity and discrimination (18). A study of illnesses in Tecumseh, Michigan also showed variation between interviewers (19). This variation appeared to diminish as the interviewers became more experienced.

In a broader context, studies conducted for the National Center for Health Statistics have added to our knowledge about interviewer effects on responses to various parts of the Health Interview Surveys (20, 21). Particular emphasis was placed on the attitudes, behavior, and background of the respondents and interviewers, and the interaction of these characteristics during the course of the interview. The procedures were a) to record the behavior of the participants during the interview using the special Behavior Observation Form in a series of verbal "snapshots"; b) to record the perceptions and general attitudes of the interviewers toward respondents and interviews; c) to obtain reactions of respondents to the interviews by later interviews with different interviewers; and d) to discuss with the interviewers their attitudes and general feelings toward their work. The findings suggested that interviews can be improved more by incorporating behavioral cues on the part of the questioner as part of the standardized interview than by changing the basic attitudes of the respondents or increasing their levels of information.

Because of the subjective nature of many of the responses evoked in assessing mood by means of a questionnaire, it seemed possible that observer variation might be particularly important in such an investigation. A community mental health assessment program in Washington County, Maryland, provided an opportunity to see if response variation was related to the interviewers, and if so, to identify factors which might have contributed to observer variation.

Materials and methods

In the summer of 1963, a nonofficial census of Washington County, Maryland,

was carried out by the Johns Hopkins Training Center for Public Health Research, Washington County Health Department, and the National Cancer Institute (22). In 1971, prior to a study of community mood in Washington County, the list of dwelling units obtained in the 1963 census was brought up to date by adding units constructed since that time. This list provided the sampling frame from which a systematic sample of 30 households was drawn weekly.

The selected households were assigned to four interviewers, each of whom was responsible for a quadrant of Washington County and the city of Hagerstown. A fifth person was employed as a back-up interviewer to take the place of the other interviewers when they were sick or on vacation. After locating the assigned household, the person to be interviewed was chosen according to random numbers in sealed envelopes. The purpose of the study and the general nature of the questions were explained to the subjects, who were told that they were free to answer as many or as few questions as they chose.

Two of the initial four interviewers resigned for personal reasons during the summer and their positions were immediately filled by the back-up interviewer and a new applicant. At the beginning, interviewers were assigned to a new quadrant of the county and city every 11 weeks. However, the rotation of assigned territories was discontinued during the latter half of the year largely because analysis of the refusal rates for the first 33 weeks showed no significant differences among quadrants. The 1584 occupied dwelling units selected during the year yielded a total of 1264 interviews, 79.8 per cent of the potential sample. Of these, 1212 were conducted by the six regular interviewers.

The questionnaire, designed by the Center for Epidemiologic Studies (CES), National Institute of Mental Health, contained a total of 435 questions under 13 different sections. Major emphasis was placed on depressed mood and feeling, and related symptomatology. The average time for completion was approximately 45 minutes. Some questions could be answered readily by the respondent but others required some explanation by the interviewer.

In addition to a comparison of the personal characteristics of the respondents allocated to each interviewer, the results of the following psycho-social tests were analyzed: a) Lubin depression adjective check list (DACL); b) CES depression scale; c) Cantril ladder—self-rating, present and future; d) Crowne-Marlowe social desirability scale; e) Langner 22-item mental health status; f) Happiness scales: Bradburn general and global; g) Aggression; h) Suicidal thoughts; i) Nervous breakdown; j) Total medications (1 week); k) Total life events (1 year); l) CES functioning scale; m) Alcohol trouble (based on Mulford) (23, 24). (Because some of the scales are composites of other scales or have been abbreviated, interested persons may obtain a copy of the questionnaire from the Training Center for Public Health Research.) All of the scales were scored so that a high numerical score indicated depression or a high level of symptom reporting. Cut-off scores were established so that the proportion of persons scoring above the cut-off point was as close to 20 per cent as possible.

Information pertaining to the interviewers was collected from several sources. The application form provided some information on their demographic characteristics. In the fall of 1972, interviewers were asked to fill out a simple form on which to indicate what kind of people they preferred to interview and what topics they considered embarrassing. Characterizations of each interviewer were also obtained from the three staff members who had had close contact with all of them (25).

Although all interviewers were white females, there was some diversity in other

characteristics. Their average age was 40.5 years, with a range from 27 to 51. All had graduated from high school, one had a college degree, and two had one or more years of college training. Their previous interview work ranged from none to experience in censuses or other health surveys. All chose to work as interviewers because of their interest in and enthusiasm for this type of work. In general, they enjoy meeting people, prefer freedom in arranging their work schedule, and do not mind daily traveling and evening calls.

In listing their preferred type of respondent, the interviewers did not indicate any particular preference on the basis of sex, race or socio-economic status except that young respondents were preferred over the elderly by five of the six interviewers. With respect to their reactions to the subject matter in the questionnaire, four of the interviewers found none of it embarrassing, one did not like to ask about church attendance, and one (interviewer C) considered four topics embarrassing: suicide, menstruation, marital happiness, and personal habits.

Although the assignment of sample households to interviewers was assigned to minimize demographic and socio-economic differences between each interviewer's group of subjects, this was only moderately successful. While there were no significant differences between groups with respect to age, sex, race and marital status, the groups did differ with respect to geographic location, household income and level of education. To a lesser extent, differences were also noted in the relationship of the respondent to the household head and church attendance habits. Because the interviewers' ratings of their respondents and the subjects' responses to the psychosocial questionnaire items might be related to the characteristics on which the groups were found to differ, statistical adjustment was desirable. This was accomplished by the use of a binary multiple regression method which allows the calculation of adjusted percentages (26, 27).

Results

The percentages of persons with "abnormal" scores (i.e., those above the arbitrary cut-off values) on the 15 psycho-social scales are shown in table 1 for each interviewer's group of respondents. Because adjustment for differences in demographic and socio-economic characteristics resulted in no important changes from the crude percentages, only the adjusted values are shown in table 1.

To judge which interviewers were associated with results that deviated from those of the others by an amount greater than that likely to have occurred by chance, a modification of the usual chi-square computation technique was used. For any specified test, the adjusted percentages were used to calculate for each interviewer the adjusted number of subjects with scores above and not above the cut-off point. These adjusted numbers were then used to calculate chi-square. Next, the procedure was repeated, omitting the most deviant value. The difference between the two chi-square values reflects the extent of variation contributed by the interviewer group which was eliminated. This procedure was repeated, successively eliminating the interviewer group with the highest contribution to chi-square, until the chi-square value for the residual group indicated a probability level of more than 0.05. For example, in the percentages of respondents with Langner 22-item scores of 3 or more, the chi-square value for all six interviewer groups was calculated to be 16.12 (d.f. 5; $p < 0.01$). The value was subsequently reduced to 3.39 (d.f. 4; $p > 0.05$) when the group interviewed by C was eliminated. In this particular instance, it is clear that all significant response variation was contributed by interviewer group C.

Significant variation between inter-

TABLE 1

Adjusted percentage of subjects with psycho-social test scores above the specified cut-off points, by interviewer groups

Psycho-social test*	Cut-off score	Interviewer groups					
		A	B	C	D	E	F
		Number of subjects					
		299	192	126	170	133	292
Lubin depression adjective check list	11+	26.3	21.8†	26.6	28.6	28.9	29.0
CES depression	16+	21.3	18.6	12.5†	17.7	18.9	17.4
Cantril ladder present	5+	25.9	28.9	15.5‡	19.6	21.2	29.2
Cantril ladder future	0+	39.5	54.4‡	41.9	41.7	44.1	49.0
Crowne-Marlowe social desirability	12+	18.9	23.2	9.5‡	22.7	18.0	15.9
Langner 22-item mental health status	3+	15.9	18.6	4.6§	16.4	19.9	19.6
Happiness, general (after Bradburn)	0	6.1	5.9	7.7†	5.5	7.0	4.1
Aggression	3+	13.3	17.5	13.1	17.9	16.3	10.9†
Happiness, global (after Bradburn)	4+	17.1	14.4	8.9‡	8.1‡	17.8	14.6
Suicidal thoughts	1+	6.5	3.4	13.7§	3.1	2.5	5.1
Nervous breakdown	1+	7.3‡	2.0‡	5.3	2.3‡	8.2‡	4.7
Total medications	6+	32.7	33.9	24.0†	27.3	31.8	32.0
Total life events	5+	17.9	15.9	8.8†	14.9	16.6	13.9
CES functioning	29+	17.2	8.4§	33.0§	24.9§	17.5	14.6
Alcohol trouble (based on Mulford)	1+	14.6	7.6	2.1§	14.3	8.6	9.7

* Tests are listed in order of appearance on questionnaire.
† Most deviant, but not significantly different from others.
‡ Significantly different at 5 per cent level.
§ Significantly different at 1 per cent level.

viewer groups was demonstrated in nine of the 15 tests. Major variation in five tests (Cantril ladder, present; Crowne-Marlowe; Langner; Suicidal thoughts; and Alcohol trouble) can be accounted for by interviewer group C. Moreover, interviewer C, as well as interviewers B and D, was associated with significant variation in responses to the CES functioning test. In addition, both interviewers C and D were associated with variation in responses to the test of Global happiness. Interviewer B was also found to have been associated with significant deviation from the other groups in responses to the Cantril ladder, future. The questions related to nervous breakdown yielded results that differed significantly for interviewers A, B, D and E.

For the remaining six psycho-social tests (Lubin DACL, CES depression, General happiness, Aggression, Total medications and Total life events), the response variation observed among interviewers could easily have occurred by chance, as indicated by the results of F tests. However, by using the same chi-square technique, one may similarly demonstrate the extent of an

individual group response from the mean score. Again, interviewer group C was found to deviate the most from the mean in four of the six tests (CES depression, General happiness, Total medications and Total life events). Interviewer groups B and F were associated with substantial response variation in Lubin DACL and Aggression tests, respectively.

In summary, the results indicated that interviewer C was consistently associated with substantial deviation from the mean scores in 11 of the 15 psycho-social tests. For nine of these 11 tests, respondents of interviewer C had lower scores than the average, and were thereby classified as being less depressed or abnormal. The exceptions were that more of her respondents said they were not generally happy and said they had suicidal thoughts; only the latter group was statistically significant. Response differences associated with other interviewers could have occurred by chance, although interviewer B, associated with major variation in two scales, is borderline in this respect.

Partial confirmation of these findings came from a subsequent study of the effects of oral contraceptives and estrogenic hormones. Among 67 respondents, interviewer C obtained a history of the use of these preparations from only 8 persons (11.9 per cent). In contrast, four other interviewers obtained positive responses from 34.9 per cent of 551 persons, with very little variation between individual interviewers (28).

In view of the response variation observed in the mental health survey, interviewer C was compared to the others with respect to work performance. She did not differ from the average in the proportion of refusals and was only slightly slower in completing an interview. However, she reported that she was able to complete her weekly assignment very rapidly, and appeared to be unusually successful in finding the desired respondent at home on the first call. Interviewer C also differed considerably from her fellows in her assessment of the subjects she interviewed. While the others felt that approximately 90 per cent of their respondents were very cooperative and 95 per cent showed a solicitous and friendly attitude, interviewer C gave these ratings to only 50 and 76 per cent, respectively. When the six interviewers were rated by their supervisors for characteristics considered to be important for successful work in this field, interviewer C was given an intermediate rating, with excessively low scores only on extroversion, confidence in her work, and coding problems.

The characteristics of the questionnaire were also related to the observed response variation. Of the nine psycho-social tests with significant variation associated with interviewers, seven were complex, with four to ten possible answers per question. Of the six tests with non-significant variation, four had only two to three possible answers per question. The nature of the subject matter also appeared to be related to interviewer-association variation. Five of the nine tests with significant response variation (Crowne-Marlowe, Langner, Suicidal thoughts, Nervous breakdown and Alcohol trouble) could be considered to deal with potentially sensitive or embarrassing topics, whereas Aggression was the only potentially sensitive test among the six not associated with significant response variation.

Two of the scales showed marked variation associated with interviewers even when the respondents of interviewer C are excluded. Greatest variation was noted in the responses to the question, "During the past week, how often did you feel that you would have a nervous breakdown?" Four categories of responses were possible. The CES functioning scale consists of questions relating to the frequency of 23 common activities during the past week, with five possible categories of answers.

Discussion

Classical thinking on the subject of interviewer effects in epidemiologic studies has usually centered on factors such as the tendency of interviewers to omit or alter certain questions or to assume the answers; the effects of probing; and the role of interviewers in explaining ambiguous or complex questions (29). The solution, in general, has focused on better training of interviewers, standardization of questionnaires and probes, and simplification of questions.

Recent investigations (20, 21) have indicated that behavioral interaction between interviewer and respondent is important in determining the amount of information obtained on an interview, with psychologic and demographic characteristics of the respondents having little effect in this respect. However, too little is known about behavioral interaction to set guidelines for the types of reinforcing behavior by interviewers that will be successful under a variety of circumstances.

A study conducted by Dohrenwend and colleagues suggested that interviewers who are embarrassed by the subject matter and who expressed a preference for certain types of respondents are likely to obtain different symptom scores than other interviewers (30). The limited findings of the present study also indicate that interviewer preference and embarrassment may be related to response variation. The one interviewer who obtained significantly different responses was one of two with more than one preferred type of respondent; in addition, she listed the highest number of embarrassing interview topics.

The unusual ability of interviewer C to find someone at home and to complete her interviews on the first home visit appeared somewhat suspicious. However, thorough checks made it highly unlikely that she failed to follow instructions for selecting a random respondent. Telephone checks of a sample of respondents and returns from questionnaires subsequently mailed to respondents showed no evidence of irregularities. Although the demographic characteristics of her subjects differed somewhat from those contacted by the other interviewers, the differences were not unusual for the area in which she worked. Although it is regrettable that periodic rotation of territories was discontinued at the time interviewer C was employed, subsequent experiences of other interviewers have shown that there were no peculiarities of the area served by interviewer C that might have accounted for her apparent success in finding respondents at home on the first visit. The most likely explanation for this is that she worked very hard on her interviewing days, making many calls but rarely recording any but the successful ones. Other evidence that she kept poor records of her visits is consistent with this explanation.

The present study was not originally designed to measure response variation associated with interviewer characteristics. Respondents were not randomly allocated to interviewers, and there was no built-in method to estimate validity of responses. Nevertheless, the results point to the following suggestions for minimizing interviewer effects. Selection of interviewers with similar characteristics and background may reduce response variation, especially in a homogeneous population. Adequate training and periodic field assessment of interviewer performance should not be neglected. Simple questions will minimize the need for interviewer assistance; the responses indicated for each question should be kept to the minimum number consistent with the possible range of answers. Periodically comparing the responses obtained by different interviewers will identify those who appear to vary markedly from their fellows. And finally, subjects allocated to each interviewer should be as representative as possible of

the entire study population in order to hold the effects of interviewer-mediated response variation reasonably constant across all population segments. In practice, this is probably most efficiently approximated by periodic rotation of interviewers through all territories.

References

1. Hyman HH, Cobb WJ, Feldman JJ, et al: Interviewing in social research. Chicago, University of Chicago Press, 1954
2. Berkson J, Magath TB, Hurn M: The error of estimate of the blood cell count as made with the hemocytometer. Am J Physiol 128:309-323, 1940
3. Gowen GH, Hall C: Dual reading for cardiovascular and other abnormalities on routine 70 mm. photofluorographic chest survey. Am J Roentgenol 79:272-278, 1958
4. Newell RR, Chamberlain WE, Rigler L: Descriptive classification of pulmonary shadows, a revelation of unreliability in the roentgenographic diagnosis of tuberculosis. Am Rev Tuberc 69:566-584, 1954
5. Power LE, Lainhart WS: Errors of diagnosis in mass radiography. Radiology 59:88-91, 1952
6. Yerushalmy J, Harkness JT, Cope JH, et al: The role of dual reading in mass radiography. Am Rev Tuberc 61:443-464, 1950
7. Armitage P, Fox W, Rose GA, et al: The variability of measurement of casual blood pressure. II. Survey experience. Clin Sci 30:337-344, 1966
8. Armitage P, Rose GA: The variability of measurement of casual blood pressure. I. A laboratory study. Clin Sci 30:325-335, 1966
9. Comstock GW: An epidemiologic study of blood pressure levels in a biracial community in the southern United States. Am J Hyg 65:271-315, 1957
10. Klimt CR, Wolff FW, Silverman C, et al: Calibration of a simplified cortisone glucose tolerance test. Diabetes 10:351-366, 1961
11. Pyke DA: Finger clubbing, validity as a physical sign. Lancet 2:352-354, 1954
12. Schilling RSF, Hughes JPW, Dingwall-Fordyce I: Disagreement between observers in an epidemiological study of respiratory disease. Br Med J 1:65-68, 1955
13. Cochrane AL, Chapman PJ, Oldham PD: Observers' errors in taking medical histories. Lancet 1:1007-1009, 1951
14. Jonathan G, Moore F, Roberts L: A discussion of technique and an analysis of errors in taking industrial histories in coal miners. Br J Ind Med 14:135-136, 1957
15. Holland WW, Ashford JR, Colley JRT, et al: A comparison of two symptoms questionnaires. Br J Prev Soc Med 20:76-96, 1966
16. Van der Lende R, Wever-Hess J, Van der Meulen GG: Investigation into observer and seasonal variation of the prevalence of respiratory symptoms at Schiermonnikoog. Uses of Epidemiology in Planning Health Services, 6th International Scientific Meetings, Proceedings. Primosten. Yugoslavia II Savremena Administracija, Belgrade. International Epidemiological Association, 1973
17. Fairbairn AS, Wood CH, Fletcher CM: Variability in answers to a questionnaire on respiratory symptoms. Br J Prev Soc Med 13:175-193, 1959
18. Fletcher CM: Some problems of diagnostic standardization using clinical methods with special reference to chronic bronchitis. *In* Epidemiology: Report on Research and Teaching. Edited by J Pemberton. London, Oxford University Press, 1963, pp 253-260
19. Wilcox KR Jr: A trial of methods for collecting household morbidity data. *In* Genetics and the Epidemiology of Chronic Diseases, USPHS Publication No. 1163. Edited by JV Neel, MW Shaw and WJ Schull. Washington: US Government Printing Office, 1965, pp 185-205
20. Cannell CF, Fowler FJ Jr, Marquis KH: The influence of interviewer and respondent psychological and behavioral variables on the reporting in household interviews. National Center for Health Statistics, Vital and Health Statistics, USPHS Publication No. 1000, series 2, No. 26. Washington: US Government Printing Office, March 1968
21. Marquis KH, Cannell CF: Effects of some experimental interviewing techniques on reporting in the health interview survey. Vital and Health Statistics. USPHS Publication no. 1000, series 2, No. 41. Washington: US Government Printing Office, May 1971
22. Comstock GW, Abbey H, Lundin FE Jr: The non-official census as a basic tool for epidemiologic observations in Washington County, Maryland. *In* The Community as an Epidemiologic Laboratory. Edited by II Kessler and ML Levin. Baltimore: Johns Hopkins Press, 1970, pp 73-97
23. Center for Epidemiologic Studies, National Institute of Mental Health. Progress report II, and Documentation of scales, Appendix 11, 1971
24. Radloff L: Revised scoring manual of CES- Community Mental Health Assessment, pilot projects: Kansas City, Missouri and Washington County, Maryland. Center for Epidemiologic Studies, National Institute of Mental Health, March, 1972
25. Lienau CC: Selection, training and performance of the National Health Survey Field Staff. Am J Hyg 34:110-132 (Section A), 1941
26. Feldstein MS: A binary variable multiple regres-

sion method of analyzing factors affecting perinatal mortality and other outcomes of pregnancy. J Roy Stat Soc 129:61-73, 1966
27. Suits D: Use of binary variables in regression equations. J Am Stat Assoc 52:548-551, 1957
28. Nomura A: personal communication, 1974
29. Hanson RH, Marks ES: Influence of the interviewer on the accuracy of survey results. J Am Stat Assoc 53:635-655, 1958
30. Dohrenwend BS, Colombotos J, Dohrenwend BP: Social distance and interviewer effects. Milbank Memorial Fund Quart 47(1, Part 2):213-226, 1969