INTERVIEWER EFFECTS IN A TELEPHONE SURVEY: A WORD TO THE WISE

Nancy Frasure-Smith

Interviewer effects, that is, systematic differences in the data collected by different interviewers, are a relatively common hazard of social science research (Bradburn, 1983; Kintz, Delprato, Mettel, Persons & Shappe, 1965; Selltiz, Wrightsman & Cook, 1976). Surprisingly, some nursing research texts pay little attention to this problem (e.g. Polit & Hungler, 1983; Wilson, 1985). Even when nursing researchers are alerted to this potential source of invalidity, when confronted with the practical difficulties of gathering data from large numbers of subjects, the possible impact of interviewer differences is often ignored (e.g. Hash, Donlea & Walljasper, 1985). What follows is the description of a study in which the interviewer variable was not considered until the analysis phase of the research, and in which interviewer differences proved to have a pervasive influence on study outcomes. It is hoped that, by presenting this result, other researchers will be reminded of the potential biases that can occur even when experienced and well-trained interviewers are involved in data collection.

The present research was conducted during the planning phases for a Quebec-wide health survey. In designing this survey the question of the stability of psychological symptoms assessed with the major survey index of mental health, the Psychiatric Symptom Inventory (PSI; Ilfeld, 1976), was of particular concern. Although itself relatively new, the 29-item PSI is a brief version of the well-known Hopkins Symptom Distress Checklist (Derogatis, Lipman, Covi, Rickles & Uhlenhuth, 1970) and was designed to assess psychoneurotic symptoms in community surveys. Initial validity studies conducted on a large community sample found that PSI symptom levels were significantly related to "having sought out professional help for emotional problems, having recently used psychoactive drugs, and interviewers' ratings of respondent's degree of tension" (Ilfeld, 1976; p.1215). In spite of this promising validity data, prior to the present study test-retest reliability had

Nancy Frasure-Smith, Ph.D. is Associate Professor in the School of Nursing and the Department of Psychiatry at McGill University, Montreal.

not been assessed in a community sample nor was there evidence of symptom stability using the telephone as the means of data collection. To meet this need we set out to examine the stability of PSI scores over a four-month period, with telephone assessment occurring on a monthly basis during that time.

Methods

Sample selection

The participation of 152 adult Quebec-born French Canadians was obtained by phoning every third French surname listed in the 1980 Lovell's city directory for streets located in a predominantly francophone, blue-collar census tract in south central Montreal. Whenever there was no answer to the first call made to a particular number, up to three additional attempts were made at varying times of the day and early evening.

Telephoners' approach

All telephoning was carried out, in French, by one or other of two French Canadian female interviewers who were in their early twenties. Both were students and had considerable experience with telephone collection of data. The use of two interviewers was necessitated by the fact that, in the course of the survey, Interviewer I began a full-time job and could not continue with the telephone work. She did, however, complete all calls for the group of respondents she had contacted, so that for each respondent the same interviewer was always involved.

Whenever anyone answered one of the calls, the interviewer worked from a standard transcript and began by introducing herself and explaining that she was working on a study of stress and health in the Montreal area. She then determined whether the respondent or anyone else in the household fit the sample requirements in terms of age (21 to 70), mother tongue (French), and place of birth (Quebec). When the respondent fit study requirements the telephoner continued by describing the amount of time the study would involve (one call of fifteen- to twenty-minutes followed by three monthly calls of about ten minutes each), and the fact that all responses would be kept confidential with the destruction of names and phone numbers at the end of the study period. All individals who agreed to take part then responded to the symptom indices followed by a brief series of background questions.

After the initial call, each subject was phoned monthly for three additional months. These calls took place as close to 30 days apart as possible, and in no case did fewer than 21 or more than 37 days elapse between calls. If a subject could not be reached within this period, the call was considered miss-

ing and the interviewer waited until the date of the next month's call to attempt to reach the subject again.

During the monthly telephone calls, the interviewers attempted to maintain as neutral a tone as possible and avoided allowing subjects to stray from the specific questions asked. The interviewers' role was that of a data gatherer rather than a support resource.

Intruments and scoring

At the time of each monthly phone call, symptom levels over the previous week were assessed using a French version of the PSI. In addition, in order to be able to compare the stability of PSI scores with the stability of a more commonly used criterion measure, the French version of the 10-item Bradburn scale (Bradburn, 1969) was also administered. It has been widely employed in studies of non-psychiatric community samples and was recently used in both English and French versions as part of the Canada Health Survey (Health and Welfare Canada, 1981). The Bradburn scale is made up of five items tapping negative affect (including things like feelings of loneliness or unhappiness) and five items tapping positive affect (including things like pride in accomplishments and excitement or interest in things). Recent research (McDowell & Praught, 1982) lends credence to the idea that positive and negative affect are independent domains among both French and English Canadians and, in spite of some minor item-specific difficulties, supports the use of the scale as a measure of emotional well-being. Besides the two measures of mental health, over the course of the study a variety of other questions were asked concerning subjects' background and demographic characteristics.

Following Ilfeld's method (1976), the PSI was scored by assigning the value of 0 to 3 to each symptom, depending on the reported frequency of that symptom. Total scores were then determined by summing these values over all 29 symptoms and converting that sum to a percentage of the highest possible sum. Thus scores could range from 0 to 100 with higher scores indicating more symptomatology.

The Bradburn scale was coded to yield three scores: positive affect, negative affect, and an affect balance score. Scores were calculated by assigning the value of 0 to 2 to each item, depending on the frequency endorsed for that item. These values were then summed and calculated as a separate percentage of the answered items for the positive and negative affect scales. The affect balance score (ABS) was obtained by subtracting the percentage score for negative affect from the percentage score for positive affect. Thus, ABS scores could range from plus 100 to minus 100 with positive scores representing a preponderance of positive affect and negative scores showing a preponderance of negative affect.

Initial acceptance rates

In order to complete the final sample of 152 subjects, a total of 1580 dialings were made to 820 different phone numbers. One-hundred and fifty numbers had been changed or had no response after three attempts. In addition, 39 individuals refused to participate before the telephoner could explain the purpose of the call. Thus, of the 820 numbers called, 631 (77%) yielded the information needed to assess their suitability for sample inclusion (screening response rate). Of those screened for eligibility, 165, or approximately 26%, did not meet study requirements in terms of age, language and birthplace. The overall acceptance rate among the 466 individuals meeting study requirements was approximately 33%. The reasons most frequently cited for not taking part in the study included not being interested, not having enough time and being bothered lately by telephone solicitation.

Table 1

Rates of refusal, drop-out and missing calls according to interviewer and sex of subject

Group	Rate of refusal[1,2]	Rate of drop-out[3,4]	Rate of missing calls[5]
Interviewer I	46.7% (N=90)	10.4% (N=48)	37.5%
Females	48.9% (n=47)	12.5% (n=24)	37.5%
Males	44.2% (n=43)	8.3% (n=24)	37.5%
Interviewer II	72.3% (N=376)	13.5% (N=104)	23.1%
Females	66.9% (n=242)	12.5% (n=80)	20.0%
Males	82.1% (n=134)	20.8% (n=24)	33.3%

[1]Ns represent the number of individuals contacted to reach the final sample sizes of 48 for Interviewer I and 104 for Interviewer II.

[2]Chi-Square, 3 df=31.01, p=.000085; 3 independent partitions: Interviewer I vs. Interviewer II, Chi-Square, 1 df=21.78, p=.00031; Interviewer I, male vs. female, Chi-Square, 1 df= .20, p=.65; Interviewer II, male vs. female, Chi-Square, 1 df=9.89, p=.017.

[3]Ns represent the number of individuals who originally agreed to participate.

[4]Chi-Square, 3 df=1.77, p=.62.

[5]Chi-Square, 3 df=5.06, p=.17

As Table 1 shows, the acceptance rates differed according to the telephone interviewer and whether the respondent was male or female. Although the interviewer with the higher overall acceptance rate (Interviewer I) had essentially the same acceptance level from men and women, Interviewer II did significantly better with women respondents than with men.

When Interviewer II's low success rate became apparent, attempts were made to improve her acceptance levels. Each interviewer worked from the same standard approach transcript, and the major difference between the two had to do with their style of delivery and voice tone. Interviewer II listened to tape recordings of the first interviewer and tried to model her delivery on these recordings. Interviewer II was also tape recorded and the first interviewer coached her to help improve her performance, but Interviewer II's acceptance rates remained unchanged. Listening to the tapes of the two interviewers it is clear that the primary difference between them was their voices; Interviewer I had a mellow, comforting voice while Interviewer II had an average speaking voice. In retrospect, when a replacement for the first interviewer was being sought, voice quality should probably have been as important a selection criterion as previous telephone experience. As will become apparent, this variable seems to have had a crucial impact on all study outcomes.

The observed acceptance rates for both interviewers are considerably lower than those reported for other phone studies involving one-shot interviewing strategies (Harlow & Hartge, 1983; Tcheng-Laroche, 1980). The reasons for this are not entirely clear, but may have involved the relatively low socioeconomic status of the respondents, the subject matter of the interview, the need for the recording of names and phone numbers, or the relatively large respondent burden associated with four monthly phone calls.

Cooperation over four months of study

In contrast to the overall low acceptance rates, cooperation over the four calls was quite good. By the time the study was completed, only 19 subjects (13%) had dropped out. As Table 1 shows, neither the interviewer involved nor the sex of the respondent was significantly related to drop out rates.

Although the great majority of subjects completed the four month monitoring period, only 72% (n=110) were reached for all four phone calls. Once again, neither the interviewer nor the sex of the subject was significantly related to missing phone calls (see Table 1). Thus, while Interviewer II was less successful than Interviewer I in recruiting subjects to the study (particularly male subjects), once subjects agreed to participate, Interviewer II was as successful as the first interviewer in maintaining their cooperation over the length of the study.

Background characteristics

The marked difference in study acceptance rates between the interviewers lead to the question of whether or not the samples obtained by them were equivalent in background characteristics. Did Interviewer II have low acceptance with all subgroups or was the difference between the interviewers most apparent among specific types of respondents? Beause we have no data on the characteristics of those who refused to participate, this issue can only be addressed indirectly by examining differences in the characteristics of the obtained samples. In addition, because of the marked difference in acceptance rates between males and females for Interviewer II, interviewer differences were examined taking the sex of the respondent into account. A number of striking contrasts emerged. The first interviewer, with her mellow, comforting voice, obtained a sample with a significantly greater proportion of blue collar males (chi-square, 1 df=4.46, p=.03), males with low levels of education (chi-square, 1 df=7.38, p=.007), housewives (chi-square, 1 df=6.26, p=.01), women with a history of psychiatric treatment (chi-square, 1 df=5.94, p=.01), and women who used psychoactive drugs (chi-square, 1 df=6.92, p=.009). No differences between the two interviewers' samples emerged for any of the other variables, including age, marital status, presence of longterm health problems, presence of longterm stresses, cigarette use, and the tendency to talk about problems with family and friends. Thus, it appears that the mellow, comforting voice of the first interviewer was perhaps less threatening, more appealing or more compelling to lower status men, non-working women and women with psychiatric problems than was the more average voice of the second interviewer. It seems likely that Interviewer II's refusals were concentrated in these groups.

Score stabiltiy

Score stability over the four monthly calls was investigated using a repeated measures analysis of variance for each score type. Estimates of the reliability of single assessments were computed using the procedures outlined by Winer (1971). Table 2 shows these estimates for the sample as a whole as well as for various subgroups based on the interviewer and sex of the subject involved.

For the sample as a whole none of the measures showed very high levels of stability. Although the PSI score was the most stable score, its reliability estimate of only .62 indicates some fluctuation over the four phone calls. In comparison, the least stable measure, the positive affect score (based on the Bradburn scale) had an overall reliability estimate of only .44, indicating that, at least in the present sample, it was probably measuring a frequently changing state rather than a more long-lived characteristic.

Table 2

Estimates of the reliability of single assessments based on repeated measures analyses of variance[1,2]

	WHOLE SAMPLE (109)	INTERVIEWER I (29)	INTERVIEWER II (80)	INTERVIEWER I Females (15)	INTERVIEWER I Males (14)	INTERVIEWER II Females (64)	INTERVIEWER II Male (16)
PSI Score	.62	.70	.52	.70	.62	.51	.53
Affect Balance Score	.57	.64	.50	.65	.55	.52	.44
Negative Affect Score	.58	.68	.41	.63	.67	.43	.43
Positive Affect Score	.44	.54	.41	.55	.52	.44	.28

[1]Includes only those subjects with responses to all four phone calls.

[2]Numbers in parentheses represent sample sizes for various subgroups.

Table 2 also shows that, as might be expected in light of the results already discussed, marked interviewer differences in stability occurred. Interviewer I, the interviewer who was more successful in recruiting subjects, also had more consistent data.

Score levels

In order to put these results into the perspective of score levels as well as degree of stability, a series of repeated measures analyses of variance were carried out. They examined the effects of interviewer and sex of respondent as well as time or call number. A number of interesting points emerge from these analyses. First, significant effects of time or call number emerged for only two scores: the PSI score and the negative affect score. In both cases multiple comparisons tests (Bonferoni t-tests; Perlmutter & Myers, 1973) revealed that scores were lowest at the third call rather than the final call. In fact, although not significant, the positive affect score and affect balance score also showed the same tendency for symptoms to increase from the third to the fourth call, almost as if subjects were anticipating the end of the study by increasing their symptom reporting.

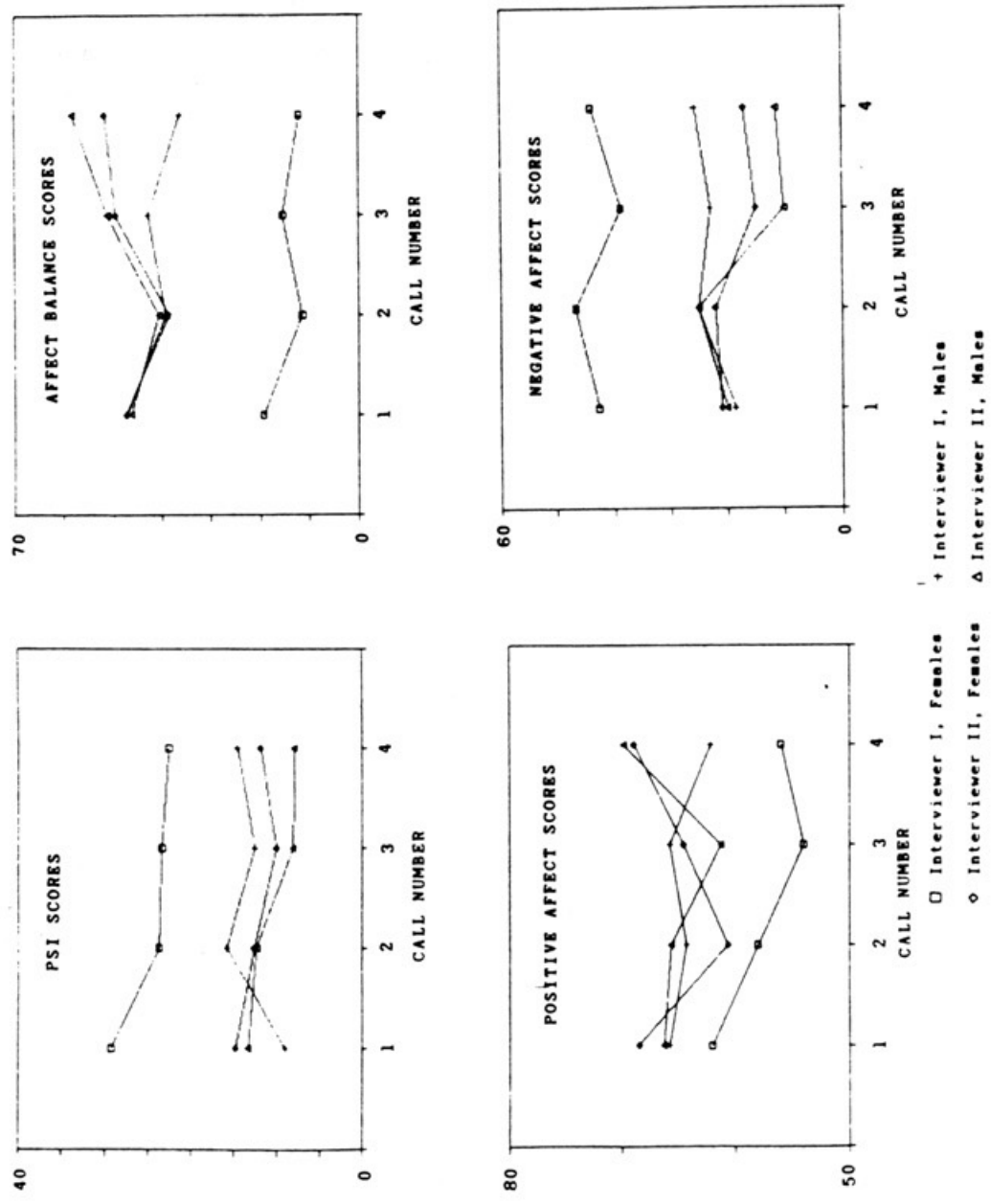

Figure 1
Mean Scores at Each Phone Call for Each Interviewer and Each Gender of Subject

Although significant interviewer and sex of subject effects occurred for all scales except positive affect, significant interactions were also present. Perhaps the simplest way to interpret these results is by examining the graphs shown in Figure 1. These figures makes one point very clear: for all scales the women subjects interviewed by Interviewer I stand out as having high symptom levels. Thus, it appears that the particular interviewer involved influenced not only initial acceptance rates and score reliability, but also the level of symptoms reported, at least among women. However, Interviewer I's subjects also included a disproportionate number of women (52.4%) who might be classified as psychiatric "cases", that is they had a history of psychiatric treatment or were currently taking psychoactive drugs. Because of this, it is not clear whether Interviewer I's impact on reliability and score levels was indirect, via the characteristics of her sample or direct and related to her influence on her subjects' symptom reporting tendencies. Unfortunately, sample sizes are too small to carry out a statistical analysis involving interviewer and caseness effects among the female respondents. However, the means for each interviewer's "female cases" and "female non-cases" were examined without testing for significance. This approach revealed that, although Interviewer I's female cases had extremely high symptom levels, there was also a tendency for the "normal" women that she interviewed to have higher symptom levels than the "normal" women monitored by Interviewer II. Thus, it appears likely that the interviewer effect on symptom reporting had both direct and indirect components.

Discussion

Although the major aim of this study was to examine the stability of scores on the PSI and Bradburn scales, perhaps the most interesting finding was the unexpected result with regard to interviewer differences. Refusal rates, sample characteristics, symptom levels and score stability all showed evidence of significant interviewer effects. Because the two interviewers worked from a standard transcript and carried out their sample solicitation in the same census tract, it is most likely that the differences between them can be attributed to their vocal characteristics, which were markedly different. Apparently this is not the first study to produce this sort of outcome. Recently, Ocksenberg, Coleman & Cannell (1986) reported the results of a study that examined the relationship between interviewer vocal characteristics and refusal rates in telephone surveys. Using a number of scales designed to assess both voice and personal characteristics, judges rated tape recordings of female telephone interviewers. Results showed that interviewers with low refusal rates tended to have higher voices, more variation in pitch, louder voices, faster speech and clearer pronunciation than interviewers with high rates of refusal. The successful interviewers were also the ones whose voices led them to be judged as the most competent and positive in their approach. Although Oksenberg and her colleagues only examined

refusal rates, and the issue of the possible influence of interviewer voices on subjects' responses was not explored, the results of the present study indicate that interviewer vocal characteristics may have major effects in telephone surveys. Further, these effects may interact with subject characteristics in complex ways. As Oksenberg and her colleagues point out, additional research into the paralinguistic aspects of telephone interviewing is needed. However, at this juncture, nursing researchers planning telephone surveys would be wise to consider the possible influence of interviewer vocal characteristics on study outcomes carefully, and to assure that data analysis is structured to examine this variable.

REFERENCES

Bradburn, N. (1969). *The structure of psychological well-being*. Chicago: Aldine.

Bradburn, N. (1983). Response effects. In P. Rossi, J. Wright, & A. Anderson (Eds). *Handbook of Survey Research*. (pp.289- 328). New York: Academic Press.

Derogatis, L., Lipman, R., Covi, L., Rickles, K., & Uhlenhuth, E. (1970). Dimensions of outpatient neurotic pathology: Comparison of a clinical and empirical assessment. *Journal of Consulting and Clinical Psychology*, 34, 164-171.

Harlow, B., & Hartge, P. (1983). Telephone household screening and interviewing. *American Journal of Epidemiology*, 117, 632-633.

Hash, V., Donlea, J., & Walljasper, D. (1985). The telephone survey: A procedure for assessing educational needs of nurses. *Nursing Research*, 34, 126-128.

Health and Welfare Canada. (1981). *The Health of Canadians: Report of the Canada Health Survey*. Ottawa: Minister of Supply and Services.

Ilfeld, F. (1976). Further validation of a psychiatric symptom index in a normal population. *Psychological Reports*, 39, 1215-1228.

Kintz, B., Delprato, J., Mettel, D., Persons, C., & Shappe, R. (1965). The experimenter effect. *Psychological Bulletin*, 63, 223-232.

McDowell, I., & Praught, E. (1982). On the measurement of happiness: An examination of the Bradburn scale in the Canada Health Survey. *American Journal of Epidemiology*, 116, 949-958.

Oksenberg, L., Coleman, L., & Cannell, C. (1986). Interviewers' voices and refusal rates in telephone surveys. *Public Opinion Quarterly*, 50, 97-111.

Perlmutter, J., & Myers, J. (1973). A comparison of two procedures for testing multiple contrasts. *Psychological Bulletin*, 79, 181-184.

Polit, D., & Hungler, B. (1983). *Nursing Research - Principles and Methods* (2nd edition). Philadelphia: J,B. Lipincott.

Selltiz, C., Wrightsman, L., & Cook, C. (1976). *Research Methods in Social Relations*. New York: Holt, Rinehart and Winston.

Tcheng-Laroche, F. (1980). *Femmes Separées ou Divorcées et Femmes Mariées: Etude comparative du niveau du stress et de l'état de santé au sein de deux groupes culturels*. Montreal: Mental Hygiene Institute.

Wilson, H. (1985). *Research in Nursing*. Menlo Park, California: Addison-Wesley.

Winer, B. (1971). *Statistical Principles in Experimental Design* (2nd edition). New York: McGraw Hill.

This research was funded by a grant from the Psychococial Research Centre of the Douglas Hospital Centre, Verdun, Quebec, and through a National Health Research Scholar Award from the National Health Research and Development Program of Canada.

The author also wishes to acknowledge the contributions of Marielle Pauzé, Lynn Clément and Joanne Kielo.

RÉSUMÉ

Effets de l'intervieweur sur un sondage téléphonique: une mise en garde

Cet article décrit l'impact des effets de l'intervieweur sur les résultats d'un sondage téléphonique visant à établir la stabilité, d'un mois à l'autre, de deux mesures de santé mentale: l'inventaire des symptômes psychiatriques (Psychiatric Symptom Inventory) et l'échelle de bien-être de Bradburn (Bradburn Scale of Well-being). Deux intervieweuses canadiennes françaises ont été contactées et ont effectué les entrevues au téléphone auprès de 152 adultes canadiens français. Les entrevues téléphoniques ont eu lieu une fois par mois pendant quatre mois. L'étude a démontré que les taux de refus, les caractéristiques d'échantillon, les évaluations de la santé mentale et la stabilité des résultats ont tous été influencés par l'intervieweuse en question. Bien que les deux intervieweuses avaient déjà une expérience des entrevues au téléphone et que les deux aient reçu une formation spéciale pour le projet, elles présentaient des différences accusées sur le plan des caractéristiques de la voix. Les chercheurs qui effectuent des études en s'appuyant sur des sondages réalisés au téléphone devraient donc accorder une attention toute particulière aux caractéris- tiques de la voix des intervieweurs et s'assurer que l'analyse des données comporte une évaluation des différences éventuelles liées à l'intervieweur.