TWO STUDIES OF INTERVIEWER VARIANCE OF SOCIO-PSYCHOLOGICAL VARIABLES

By: Leslie Kish and Carol W. Slater
Survey Research Center, University of Michigan

Introduction

We report results obtained in two surveys in which respondents were randomized among interviewers to permit the valid estimation of the interviewer variance as a component in survey errors. In each study, done by the Survey Research Center of the University of Michigan, the blue-collar workers of a plant were asked many socio-psychological questions about their jobs, and company and such.

We suspect that few people worry at all about the interviewer variance. They, however, are apt to fear that for "vague" psychological and attitudinal questions the effects must indeed be large. Our results may hold at least one surprise for everybody.

On the one hand, the interviewer effects are not very great: they compare well with effects on "factual" items, and, because of this we were unable to separate different classes of items - the "soft" psychological items from the "hard" factual items. On the other hand, even these small or moderate effects on individual interviews can have important effects on the sample means. As a final dramatic effect, a happy ending: even great effects on the means of the entire sample are reduced for subclasses and the effects usually seem to disappear completely from the comparisons of subclasses.

Here we present a summary of our findings; details and references to related literature will appear in an article already submitted for publication.

In the First Study in 1948, at a large unionized auto plant in the Midwest, we selected with equal probability a stratified random sample of individual employees, of whom 462 gave interviews. The names and addresses of the selected employees were typed on cards which were then shuffled and assigned randomly to interviewers at the beginning of each day. The interviews were taken in the respondents' homes and lasted an average of an hour and a half. Open-ended questions were used to gather information about attitudes towards foremen, stewards, the union, higher management and various aspects of their jobs. The 20 interviewers were selected, screened and hired specifically for this study. All had had some previous experience in interviewing, not necessarily in survey work. A week of training was carried out before the study began and was augmented, as needed, by individual supervision and group sessions.

For the Second Study, we selected in 1958 with equal probability a stratified random sample of individual employees, with a final n = 489. The interviews were conducted, in 1958, in offices provided by the company, and lasted an hour on the average. After the interview, each respondent was also asked to fill out a paper-and-pencil questionnaire in the presence of the interviewer, which took roughly an additional three-quarters of an hour. From the list of respondents available during each week, random assignments were allocated to the interviewers working during that period. Completely open-ended items were few: almost all questions included in the written questionnaire and many of those used in the interview involved asking the respondent to choose from a prepared and pretested list the alternative coming closest to his own viewpoint. The nine interviewers were members of the Center's field staff with several years of interviewing experience.

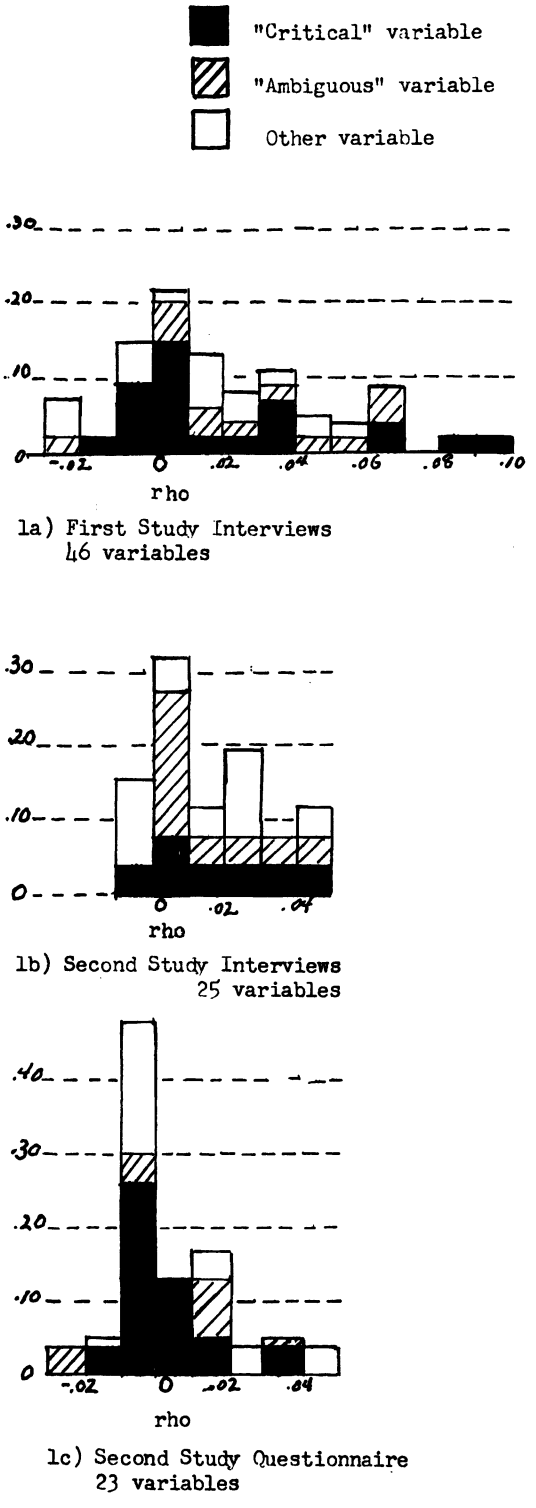
The Measurement of Interviewer Variance

Besides sampling errors proper--those arising in selection or estimation procedures--survey results are also affected by errors which occur in the course of the observation (measurement), recording and processing of the data. These errors fall into two broad types, having very different effects on the summary results (such as means or totals) of a survey. The first includes the "biases" or "systematic errors" imposed by the "essential survey conditions": the average or "expected" deviations of sample estimators from their estimands, the population values. These, although important, were not the subject of our research.

The second type consists of variable errors: those not fixed by the "essential survey conditions." Some variable errors are uncorrelated among the elements, and, unless replicate measurements are taken on the respondents, these cannot be distinguished from sampling error among respondents. We are not here concerned with them and generally they can be regarded as random errors which increase the variance of estimators with contributions which enter automatically into the estimate of the variance. Some other variable errors, however, involve the correlated effect that each interviewer's bias can impose upon the respondents (the elements) making up his workload. Insofar as the individual interviewers have different average effects on their workloads, this "interviewer variance" contributes to the variance of the sample mean. This contribution of the interviewer variance to the sampling variance is our present concern. The contribution, as we shall see, can be large and its neglect can lead to serious underestimation of the total survey variation.

Our model assumes the random selection of a sample of a interviewers from a large pool of potential interviewers, that pool defined by the "essential survey conditions." Each interviewer has an individual average "interviewer bias" on the responses in his workload; we estimate the effect of a "random sample" of these biases on the variance of the sample mean. This effect is expressed as an "interviewer variance" which decreases in proportion to the number (a) of inter-

CHART 1 - Three Distributions of Relative Frequencies of Rho's for Different Variables



1a) First Study Interviews
46 variables

1b) Second Study Interviews
25 variables

1c) Second Study Questionnaire
23 variables

viewers. Its contribution to the variance of sample means (s_a^2/a) resembles other variance terms, being directly proportional to the variance per interviewer and inversely to the number of interviewers. This increase in the variance may be substantial; failing to take it into account (as when estimating the variance simply by s^2/n) results in neglecting a potentially important source of variation actually present in the design, introduced by the sampling of interviewer's biases.

The interviewer variance s_a^2 should be viewed as a component of the total variance, denoted as

$$s^2 = s_b^2 + s_a^2$$

where s_b^2 is the variance without any interviewer effect, and all three terms are measured per element. It is convenient to take the interviewer component s_a^2 relative to the total variance, and to denote this ratio by the ratio of homogeneity, often called the coefficient of intraclass correlation:

$$roh = s_a^2/s^2 = \frac{s_a^2}{s_a^2 + s_b^2}$$

The individual roh's are subject to very great variabilities; the values of s_a^2 are computed with 9 degrees of freedom in the First Study and 19 in the Second Study. As a rough guide we consider the values of the First Study subject to coefficients of variation of 0.5 and those of the Second Study to about 0.3. Nevertheless, the results are useful, particularly when considered in the aggregate over many items.

Primary Results and Implications

How are these values useful in planning surveys? First, they show that it is feasible to obtain responses with rather low interviewer effects on what appear to be ambiguous and emotionally loaded attitudinal items, if the interviewers are carefully selected and well-trained. The low values of roh on these items speak well for the prospects of obtaining attitudinal, socio-psychological data of this kind with reasonable reliability. The variability for these attitudinal interview items appear to be generally not much, if any, higher than responses to "factual" items obtained in a good Census--expect probably for the simplest items like age and sex. They compare favorably with some other results relating to "factual" items.

The primary results appear on Chart 1; the First Study in 1a, the interviews and questionnaires of the Second Study in 1b and 1c, respectively. Each of these presents a distribution of the relative frequencies (percentages) of occurence of rohs in size classes of .01. (The total height of each class is divided into three to separate "critical", "ambiguous" and other items.)

Second, this kind of analysis can distinguish

items for which the interviewer variances appear unexpectedly high, and by so doing, lead to corrective actions either through better training or by changing the survey operations. Extension of this kind of analysis may also be used to separate interviewers who make undue contributions to the variances.

Third, we can distinguish in the three tables concomitants of different interviewing situations. The results of the First Study (1a) came from newly hired and trained interviewers taking open-ended interviews; the roh's range, in the main, from zero to .07, with an average of .02 or .03. In the Second Study we see expert interviewers taking a more structured interview (1b); the roh's vary mostly from zero to .04, with an average of .01 to .02. For written questionnaires we find (1c) that the a priori hypothesis of zero effect is generally acceptable (with the exception of three puzzling items).

Fourth, our results indicate the difficulties involved in making judgments beforehand about the degree of interviewer variance associated with what may seem a priori to be different kinds of items. In each of the three parts of Chart 1 the areas corresponding to "critical", to "ambiguous" and to "other" items do not appear to have very distinct distributions. Even informed intuition, it seems, needs considerably more conceptual and empirical tools than are now available to evaluate the relative susceptability of survey items to interviewer bias.

Fifth, we find that interviewer variance, although it appears small, definitely exists. Furthermore, it can exert important influence on the total variability of survey results, since even a small roh, when multiplied by moderate or large interviewer workloads, can have large effects. This effect on the variance is about $[1 + \text{roh}(\frac{n}{a}-1)]$. Let us consider an increase in the variance by a factor of 1.5 as "serious" and by 2 as "critical"; these correspond to increases in the standard errors of $\sqrt{1.5} = 1.22$ and $\sqrt{2} = 1.41$. With $n/a = 22$ in the First Study, roh becomes serious at .025 and critical at .045 categories which include 16 and 8 items respectively. In the Second Study, with $n/a = 52$, roh = .01 is serious and roh = .02 is critical, thus including 13 and 10 items respectively. In the case of element sampling; these effects can and should be included in the variance by computing the interviewer's load as if it were a "cluster." (In the case of actually clustered samples, where the interviewer is confined to a single primary selection, such as a county in a national sample, the usual computation automatically includes this effect.)

Sixth, analysis of this type makes it possible to include interviewer effects in considering the economic aspects of survey designs. If the ratio of the cost of hiring and training an interviewer to the cost of a single interview is $\frac{C_a}{C_b}$, then the most economical plan - least total variance ($s_a^2/a + s_b^2/n$) - results from the optimum workload size of $\frac{n}{a} = \sqrt{\frac{C_a}{C_b}\frac{s_b^2}{s_a^2}} = \sqrt{\frac{C_a}{C_b}\frac{1-\text{roh}}{\text{roh}}}$.

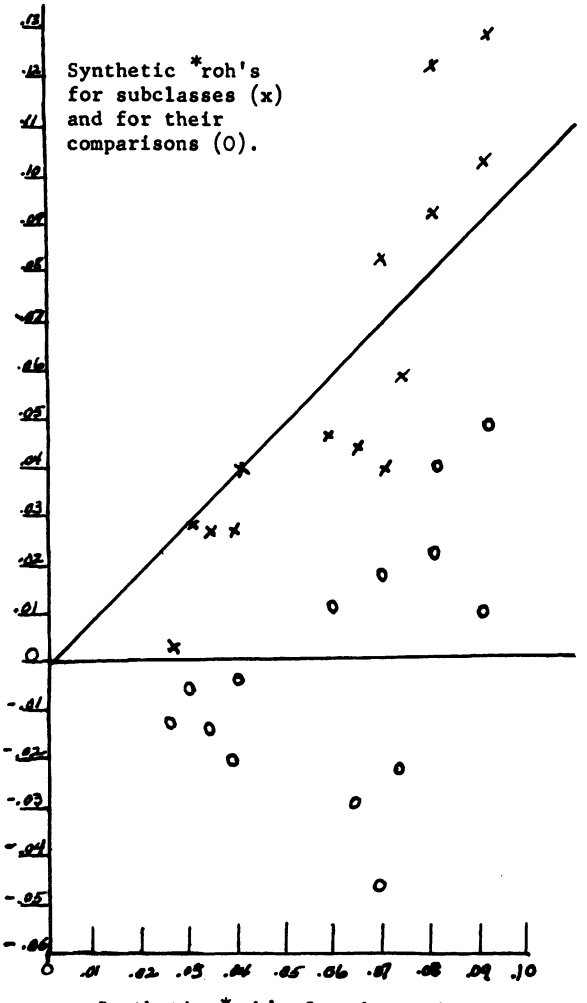
For example, if it costs ₹180 to train an interviewer and ₹10 to take an interview, then $\frac{C_a}{C_b} = 18$. For roh's of .02 this gives an optimum workload per interviewer of $n/a = 30$. The actual workloads in our two studies were in this neighborhood.

Effects on Subclasses and Their Comparisons

Current models of response errors deal mostly with the effects on the mean for the entire sample, but applying the model and methods to the means of subclasses is straightforward.

The data support our hypothesis that the effects of interviewer variance on the variances

Chart 2 - The Effects of Interviewer Variability on Subclasses (x) and on Their Comparisons (0) Plotted Against the Effects on the Entire Sample. (The effects are measured as ratios to the total variance per interview - as synthetic equivalents of roh's.)



of subclass means tend to decrease in the same proportion as the average workloads of the subclasses per interviewer decrease. The effect on the variance is approximately $[1 + roh(n^*/a-1)]$, where n^* is the sample size of the subclass. This effect decreases if roh remains constant, where roh expresses the interviewer contribution per element. That roh remains fairly constant for the subclasses is evidenced by the proximity of the values marked by x to the 45 degree line on Chart 2; this line denotes equality for the roh's of the subclasses and of the entire sample. The x points mark the values of roh in subclasses against the roh for the entire sample for the same variable. Actually the ordinates denote the average of the roh's for two subclasses into which the entire sample was divided.

We also investigated the effects of interviewer variance on the comparisons of pairs of subclass means. These are even more important, research workers often say, than the estimation of individual means. Because of the considerable effort required we had to limit the extent of this investigation. For strategic reasons we chose for investigating both the subclasses and their comparisons seven of the most critical variables from the two studies: those for which the effect on the means were greatest. For each variable we used two different ways of forming subclasses and this gives rise to the fourteen comparisons marked O on Chart 2 (as well as the fourteen subclass averages marked x).

The results show that the effect of interviewer variance on comparisons between subclass means is reduced drastically to the neighborhood of zero. This important result seems to hold roughly and on the average in our investigation. As evidence, note on Chart 2 that the O marks for interviewer effects on comparisons fluctuate around the horizontal line denoting zero effect. These come from plotting the effects per element of the comparisons against those of the entire sample. These data show a great deal of fluctuation, the causes of which should be sought in later investigations; nevertheless, the tentative working hypothesis of zero average effects appears to be a good working hypothesis on the average, and better than any alternative we could form. This should apply also to comparisons of any two (or more) samples which have been randomized over the same set of interviewers. Important examples arise from the comparisons of periodic samples assigned to the same set of interviewers; such comparisons of periodic surveys should tend also to be free of the effects of the interviewer variance that affects a single sample. (Similar results were obtained on the very different data, with very high initial roh's.)

This result gains added significance in combination with the likelihood that the systematic biases of comparisons are often also less than the biases of the individual means. In other words, if the interviewers' biases affect the subclasses equally (corresponding to lack of "interaction" between interview bias and subclass) then both the systematic bias and the interviewer variance tend to disappear from the comparisons of subclasses.

Some Remarks on Research Strategy

Research on interviewer variability may be designed to different degrees of symmetry and completeness. A very complete design might call for simple random selection of equal workloads; the effects of other sources of errors, especially coding error, would be included in a symmetrical, clean (orthogonal) design; the questions could be chosen to test various hypotheses about them. Our studies lack these virtues.

Our randomization procedures were designed to minimize costs and interference with field operations. We sacrificed chiefly: (a) equal size workloads which would have resulted in somewhat simpler computations and more efficient estimates; (b) eliminating the complications arising because the randomized set (the workload for a day in the First Study and for a week in the Second Study) is difficult to treat exactly in the analysis; (c) and the possibility of separating the components of the variance due to coding variability by randomizing coders in a neat design. Perhaps we most regret lacking the means and persuasiveness to achieve the last of these three improvements--which a modest disposal of means could have brought. In defense we plead that the choice was between little or nothing--as it often is. The procedure for assigning interviewers to coders does not depart enough from random to interfere seriously with our analysis: the distribution of coders against interviewers was checked and found about as even as a random assignment would have made it.

With relatively modest extra means it is possible to get a little closer than we did to a more symmetrical and complete design. Nevertheless, we are convinced of the desirability and economy of allocating near the lower end of the scale the limited resources available for research in interviewer variability. This is not merely post hoc justification for our research, but a belief based on the expectation that a few--because expensive--"crucial experiments" will not yield definitive evidence about a small set of "basic parameters"-- because that small set does not exist. It is more likely that interviewer errors differ greatly for various characteristics, populations, designs and resources--this last including questionnaires, nature and training of interviewers, etc. Therefore, knowledge about this source of variation, as with sampling variability, can be accumulated only from a great deal of empirical work spread over the length and breadth of survey work. This implies, together with the necessarily limited total means for this kind of research, that most research in this area must be done at marginal cost, as appendages to the main aims and designs of surveys.

Therefore, general strategy should call for many investigations of modest scope and that these be widely communicated.

APPENDIX

The response from the j-th individual to the i-th interviewer is expressed as $y_{ij} = y'_{ij} + A_i$, where A_i is the average "effect" of the i-th interviewer. Any constant (or "systematic") biases of the interviewers are not distinguished and we assume that the sum of the interviewer effects is zero for the population of A interviewers, from which the actual a interviewers are a random sample.

Each response is viewed as composed of two components, the sampling variance of the individual response and the component due to the variable interviewer bias, or interviewer variance:

$$s^2 = s_b^2 + s_a^2 \quad .$$

The "ratio of homogeneity" is

$$\text{roh} = s_a^2 / s^2 = s_a^2 / (s_a^2 + s_b^2). \qquad (1)$$

Assuming that of the n respondents n_i were assigned with simple random sampling to the i-th interviewer, we have in terms of the usual "anova" table for computations:

Source of Variation	Degrees of Freedom	Sum of Squares (SS)	Mean Square	Components of the Mean Squares
Among interviewers	a-1	$\sum^a y_i^2/n_i - y^2/n$	$V_a = \frac{SS(a)}{a-1}$	$s_b^2 + ks_a^2$
Within interviewers	n-a	$\sum^a \sum^{n_i} y_{ij}^2 - \sum^a y_i^2/n_i$	$V_b = \frac{SS(b)}{n-a}$	s_b^2

Here $n = \sum^a n_i$, $y = \sum_i^a y_i$ and $y_i = \sum_i^{n_i} y_{ij}$ and $s_a^2 = (V_a - V_b)/k$ where

$$k = \sum^a n_i^2 \frac{1/n_i - 1/n}{a-1} = \frac{n}{a-1} - \frac{\sum^a n_i^2}{n(a-1)} = \frac{n}{a} - \frac{1}{n}\left[\frac{1}{a-1}\sum^a (n_i - \frac{n}{a})^2\right] .$$

To measure the effects on the differences between the means of two subclasses we had to improvise approximate methods. To compare with the preceding we began by computing the "effect" of interviewer variance as the ratio of the actual variance to the simple random variance for the entire sample:

$$e(\bar{y}) = \frac{\text{var}(\bar{y})}{s^2/n} \quad \text{where} \quad s^2 = \frac{1}{n-1}\sum_i^a \sum_j^n (y_{ij} - \bar{y})^2$$

and $$\text{var}(\bar{y}) = \frac{1}{n^2}\frac{a}{a-1}\left[\sum^a y_i^2 + \bar{y}^2 \sum^a n_i^2 - 2\bar{y}\sum^a y_i n_i\right].$$

This last is the variance of the ratio estimator $\bar{y} = y/n$ of a randomly selected clusters. The computed effect on the variance is then equated with $[1 + {}^*\text{roh}(n/a - 1)]$ and this yields the synthetic ${}^*\text{roh} = [e(\bar{y}) - 1]/(n/a - 1)$. (2)

Column 4.

We expected and found that formulas (1) and (2) gave very similar results. We then computed using (2) the values of $e(\bar{y}_1)$ and $e(\bar{y}_2)$, the effects on the variances of the means of two subclasses. We computed actually $e(\bar{y}_1) = \text{var}(\bar{y}_1)/(s^2/n_1)$, and similarly for the other subclass; that is, we did not bother to compute separately s_1^2 and s_2^2 because they would not have differed enough from s^2 to make that extra labor worthwhile.

The synthetic *roh's for the two subclasses were averaged and these were plotted as x in Chart 2 against the roh's for the entire sample.

The variance of the difference $(\bar{y}_1 - \bar{y}_2)$ was computed, taking into account the correlations within the workloads of the a interviewers, as:

$$\text{var}(\bar{y}_1 - \bar{y}) = \text{var}(\bar{y}_1) + \text{var}(\bar{y}_2) - 2\,\text{cov}(\bar{y}_1, \bar{y}_2).$$

As with the variances, the covariance is computed for the ratio estimator of a randomly selected clusters:

$$\text{cov}(\bar{y}_1, \bar{y}_2) = \frac{1}{n_1 n_2}\frac{a}{a-1}\left[\sum^a y_{1i}y_{2i} + \bar{y}_1\bar{y}_2\sum^a n_{1i}n_{2i} - \bar{y}_1\sum^a y_{2i}n_{1i} - \bar{y}_2\sum^a y_{1i}n_{2i}\right]$$

From these we computed the effects on the difference

$$e(\bar{y}_1 - \bar{y}_2) = \frac{\text{var}(y_1 - y_2)}{s^2(1/n_1 + 1/n_2)}$$

Finally, we computed the "synthetic *roh" as

$${}^*\text{roh} = \left[\frac{\text{var}(\bar{y}_1 - \bar{y}_2)}{s^2} - \frac{1}{n_1} - \frac{1}{n_2}\right] \div \left[\frac{2}{a} - \frac{1}{n_1} - \frac{1}{n_2}\right].$$

These values appear as the 0 points on Chart 2 plotted against the values of roh for the entire sample.