SOME SUGGESTED METHODS OF DETECTING INTERVIEWER DIFFERENCES IN SAMPLE SURVEYS[1]

BU-44-M J. E. Dowd June, 1953

Introduction and Statement of the Problem

In recent years much attention has been drawn to the problem of non-sampling errors in surveys. While there are now adequate methods available for controlling, or at least measuring, sampling errors in the usual probability survey, little is known about the sources of non-sampling error, or what their effects are upon sample estimates. One can never be quite sure that his sample averages do not contain some element of bias, or that his sample variances are not inflated by errors other than what his formulae for standard error allow.

Non-sampling errors, as their name implies, are independent of the sampling procedure and may enter into the survey at any point from the designation of the objectives to the conclusions drawn from the results (1, 1944). Most concern, however, has been directed toward those errors which arise in the collection of the data from the respondents. Much has been written on the problem of non-response due to both those not at home when the interviewer calls and those who refuse to be interviewed. Methods have been devised for including these people in the sample or for getting unbiased estimates from the sample when these people are not included. Recently, too, concern has been expressed over those errors occurring during the interviewing situation. These errors may arise from the interaction of the interviewer and the respondent or may be due to errors on the part of the interviewer or respondent alone. Some of the more important causes will be mentioned in greater detail in a later section of the paper.

[1] Report of research prepared in connection with the National Analysts, Inc. - A. J. King Grant.

Interviewer errors may be thought of as arising from two fundamental sources:

1. Differential interviewer effects which are a reflection of the individual interviewer's idiosyncrasies upon the recorded response.

2. An overall bias which will affect the results of all interviews in the same manner.

For example, it has been noted that when eye estimates of crops have been made by two different groups of enumerators on the same crop there is individual variation in the estimates of different enumerators and also a bias common to each group depending upon what training methods were used for each group and under what auspices the groups are working. It should be noted that while a group of interviewers may be completely homogeneous in the results they elicit; i.e., there is no differential response within each group, the group as a whole may be biasing the results.

It is important to distinguish between these two sources of error, for when dealing with a single group of interviewers, while we shall be able to test whether differential interviewer effects exist we cannot test whether our interviewing group as a whole is biased unless we have outside means of verifying the estimates. For example, we could test the validity of such factual data as date of birth, ownership of a telephone, driver's license, etc. by checking with official records. It would be much harder, though, to try to validate such intangibles as opinion or attitude structure since a major problem would arise in attempting to define an individual's "true" opinion or attitude.

While individual interviewer effects may be self cancelling (there is no guarantee that this will be so) the variation amongst interviewers may contribute a greater portion to the total error of the survey than will sampling variation.

It may also be that one or two of the interviewing staff may be getting results which are not in line with the results of the rest of the interviewers. Where a small interviewing staff is being used, the discrepant results of these one or two interviewers may have an unduly large effect upon the final estimates. It will be of interest, then, to have some estimate as to the interviewer's contribution to the total variation along with a check as to whether certain interviewers are obtaining discrepant results. No attempt will be made to suggest methods for correcting any observed differences in results nor will there be any attempt to hypothesize the causes of observed differences. These considerations are outside the scope of this paper and depend to some extent upon the practical situations arising out of any particular survey. It shall be left to persons who are familiar with the content of the survey in question and the makeup of the interviewing staff to decide what action (if any) should be taken when interviewer effects are shown to be present.

We shall further confine ourselves to testing for differences when the following restrictions are imposed:

1. Actual survey conditions are in operation.
2. Respondents are chosen by a random device and not by the interviewer.

The detection methods suggested in this paper have been formulated with the aim of being carried out as an integral part of the survey design. The methods should be subordinated to the usual survey administrative considerations and should be incorporated into the survey in such a manner as to keep added costs at a minimum. Experiments have been set up to test whether interviewer differences exist either between certain groups of interviewers (experienced vs. inexperienced, training method A vs. training method B, etc.) or between individual interviewers within a group. For the most part, however, the experimental results have been the predominant objective and the conditions under which these experiments were carried out called for a compromise from the usual survey techniques in order to assure valid comparisons of the experimental results.

It is desirable to eliminate from this discussion the so-called "judgment" and "quota" sampling techniques since when they are used there is no estimate of sampling error available and there is also an additional potential source of bias present due to the fact that the interviewer is allowed to choose the individuals he is to interview. Certain interviewers might be prone to choose certain types of respondents within their quota groups, e.g., those people who seem easiest to approach or those people the interviewer thinks will give the "best" kind of interview. The obvious solution to this problem of added bias present in the quota type survey is to use a probability sample.

Review of the Literature

The overall problem of interview bias has been approached in a fundamental manner by members of the National Opinion Research Center in a research program sponsored by the National Research Council and the Social Science Research Council through their Joint Committee on the Measurement of Opinion, Attitudes, and Consumer Wants. The results of the project have not yet been published in a unified report but in an interim report (2, 1949) H. Hyman has outlined the program of research followed at that date:

1. Examination of existing surveys for variables related to effects, magnitudes of such effects, and the process by which they occur.

2. Intensive interviews of interviewers concerning their experiences.

3. Intensive interviews with a small sample of respondents.

4. Measurement of interviewer differences under national survey conditions where interviewers were given equivalent assignments using both homogeneous and heterogeneous groups of interviewers.

Although a complete report of findings has not yet been published, various aspects of the project have been dealt with in articles by members of the N. O. R. C. staff appearing for the most part in Public Opinion Quarterly and the International Journal of Opinion and Attitude Research. Hypotheses as to the presence of interviewer effect have been tested with regard to the following situations:

1. Different types of questions, e.g., dichotomous with alternatives explicitly stated, categorical eliciting many answers, etc., were compared for the purpose of testing whether certain types of questions allowed interviewer effects to operate through them. It was found that four types of questions predominated in allowing effects to operate through them (3, 1947).

2. In area sampling a study of interviewer performance was made (4, 1949) in which it was shown that interviewers make conscious or unconscious errors in listing dwelling units. There is also a possibility of error in the selection of individuals within a dwelling unit.

3. A study of the influences of sub-questions on interviewer performance (5, 1949) showed that there is no marked tendency for interviewers to classify answers into categories which will avoid further questioning by means of sub-questions. It was found, however, that those interviewers who favored a question themselves received more replies to it than those who did not favor it.

4. A study was made to show effects by classifying answers into pre-coded boxes. It was found that interviewer attitudes and expectations have little effect upon the classification of response (6, 1949).

5. The hypothesis that interviewer expectations as to the organized structure of the respondents' attitudes distort survey results through errors in recording was tested (7, 1950). The hypothesis was not rejected and it was postulated that interviewer expectations could operate in the following manner:

a. Role expectations: The interviewer views the respondent as a member of some group and the response is stereotyped accordingly when there is any room for the interviewer to interpret ambiguous answers.

b. Attitude structure expectations: The interviewer assumes that the attitudes of a respondent must in some way be consistent, e.g., those who are pro-Democrat will be pro-labor.

c. Probability structure expectations: The interviewer enters the interview situation with certain concepts of opinion distribution, e.g., nearly everyone favors N. A. T. O.; few doctors favor socialized medicine. Besides the tendency to interpret ambiguous responses in light of the interviewer's expectations it is also thought that the interviewer will also tend to oversimplify the responses.

6. A field study with a large number of interviewers was carried out in the city of Denver to test whether or not interviewer opinions had any influence on response (8, 1951). In this study an attempt was made to insure comparability of results by dividing the city into sectors, each of which was composed of a representative group of census tracts. A group of interviewers would then interview in all census tracts in one sector. Comparisons were first made for each group within a sector and then results for all sectors were pooled allowing comparisons for the whole staff of interviewers. Negative results were obtained inasmuch as the hypothesis of interviewer effect was rejected.

Studies on biases and on differences among samples in visual methods of estimating various characteristics of field plots have been reported by a number of authors (9, 1934; 10, 1935; 11, 1936; 12, 1940). In those experiments that were set up in such a way as to allow comparisons among observers (for example, 9, 11) it was shown that there is also a variation in amount of bias obtained by different observers. Yates and Watson (9, 1934) report the use of a 10x10 latin square to test variation in bias when it was necessary to determine by means of an eye estimate the number of shoots of a plant that had reached a certain height. The results were checked for bias by digging up all plants on the experimental plots. In the experiment reported ten experimental plots were each examined by ten trained observers. The plots were checked in a certain order both to insure that two observers were not examining the same plot at the same time and to determine whether the order in which the plots were checked had any influence on the magnitude of the bias obtained. The rows of the latin square corresponded to the observers, the columns to the plots, and the treatments to the order of examination. It was found that all observers tended to underestimate the number of shoots and that there was also a significant difference in the magnitude of the biases of different observers. The amount of bias also differed significantly according to the order in which the plots were examined. This was thought to be due to the fact that those who first examined a plot disentangled and separated the shoots of the plants thus making a more accurate estimate possible by those who later examined the plots.

Studies by workers at Iowa State College (13, 1940; 14, 1942; 15, 1946; and 16, 1946) and in India (17, 1947) have shown that the yield of a crop is overestimated if the sampler is allowed to use any judgment whatsoever in the placement of the sampling frame. In placing the sampling frame the question always arises as to which plants to include and which to exclude. The tendency is to include too many. There is also a tendency to choose a better than average spot in the field in which to throw the hoop (13, 14, 15). This tendency may be overcome by choosing a random spot in the field over which to place the hoop (14). When a crop looks poor the observers must also use judgment as to whether or not the field will be harvested. Unless some standard criteria are agreed upon for deciding whether or not a crop will be classed as harvestable there will be discrepancies in the results of different enumerators.

Another example in which a survey was designed so that it is possible to get "unconfounded" estimates of the effects of two or more groups of interviewers is described by Durbin and Stuart (18, 1951). A factorial experiment is used to test whether differences in the rate of response elicited exist between experienced and inexperienced interviewers. The design used is a 4x3x3x3x2 factorial replicated seven times, the factors being as follows:

4 ages of subject

3 questionnaires

3 groups of interviewers
2 experienced
1 inexperienced

3 city boroughs of districts

2 sex of subject

A total of 1512 interviews were attempted within this experiment, 504 by each interviewer group. It was possible to carry out the experiment testing simultaneously whether the experience of the interviewer, type of subject matter in the questionnaire, district in which the survey was carried out, age of respondent, and sex of respondent, had any effect upon the response rate obtained. It should be noted that the use of such a layout does not enable us to detect whether or not differences exist between individual interviewers but only between interviewer groups, i.e., if we have found that the groups differ we cannot identify the particular interviewers in the group that are causing the discrepant results of the group.

It was found in this experiment that the inexperienced interviewers did obtain a significantly lower response rate but since the inexperienced interviewers were all university students while the experienced interviewers were members of two commercial survey firms no conclusions could be reached as to whether the lower response rate was due to the inexperience of the group or to some other characteristic which might be ascribable to the student group.

As opposed to those surveys in which an experimental layout or design is used to allow valid comparisons of interviewer results and whose main objective is the testing of some hypothesis about certain groups of interviewers or individual interviewers, there have been developed methods which, when incorporated into a survey, will allow a comparison of the results of individual interviewers or groups of interviewers. These comparisons are not the main objectives of the survey but are an intrinsic part of the actual survey design.

Since 1938, workers at Iowa State College (see 13, 1940; 14, 1942, and 15, 1946) under the direction of A. J. King were conducting route-sample surveys to estimate crop acreage and production of field crops. Two samplers were assigned to each car. Upon reaching the designated field each sampler took one of the two designated samples from each field. By identifying the samples it was possible to compare samplers. For example, in the soybean survey reported by Houseman et al. (15, 1946) the following analysis of variance between samplers was obtained:

Source of variation	Samplers A and B		Samplers A and C	
	df	ms	df	ms
Samplers	1	0.24	1	188.36
Fields	45	118.20	19	83.83
Samplers x Fields	45	73.72	19	79.77

The results from samplers A and B agree very closely. However, it is known that the data obtained were not the result of collusion between the samplers but a chance event. The difference between samplers A and C is not unusually large. King and Jebe (13, 1940) also mention a discrepancy in the results of two observers although no data is given on the discrepancy.

P. C. Mahalanobis has played a large part in the development of these methods and has made use of them in much of his survey work. One method which he describes (19, 1946) is most commonly used by him and his associates. Within each stratum in which the sampling units or clusters of sampling units are located at random the sampling units are given serial numbers as they are located. Two or more independent interpenetrating samples are made from those units which have even-ending serial numbers and those that have odd-ending serial numbers. These samples are then enumerated by different interviewers or groups of interviewers, thus supplying different estimates of error. This method allows one to make comparisons within strata but not for different interviewers working in different strata. This technique also requires that both enumerators cover the whole stratum or segment rather than each one working within a more compact portion of the segment. This means that travel costs will be higher for all interviewers and a greater length of time will most likely be required for the completion of the survey. This type of design fits itself best to a survey situation in which a small group of interviewers travel together from stratum to stratum, sharing the interviewing within each stratum, e.g., a sociological survey carried out by a university where the interviewing staff might consist of two or three trained sociologists travelling in the same automobile within a state.

Hochstim and Stock (20, 1951) have reported a method for "measuring interviewer variability". They assume the following additive model:

$$y_{ij} = x_i + a_j + \epsilon_{ij}$$

where y_{ij} = an observation on the ith respondent by the jth interviewer,

x_i = the "true" value of the characteristic measured on the ith respondent,

a_j = the effect of the jth interviewer in repeated observations on the whole population, and

ϵ_{ij} = a random error.

Assuming that k interviewers are assigned at random to n sampling units they use the analysis of variance technique for a completely randomized design to estimate the component of sampling variation due to interviewers. The analysis of variance is:

Source of variation	df	ms	ms an estimate of
Between interviewers	k-1	B	$\sigma^2 + k'\sigma_B^2$
Between respondents	n-k	A	σ^2
Within interviewers total	n-1		

where $k' = \frac{1}{k-1}\left(n - \sum_{j=1}^{k} \frac{n_j^2}{n}\right)$ where each interviewer interviews n_j respondents and $\sum_{j=1}^{k} n_j = n$,

σ^2 = the common variance of the ϵ_{ij}'s

and $\sigma_B^{\,2}$ = the common variance of the a_j's.

If we assume that the a_j's and ϵ_{ij}'s are independently and normally distributed the ratio B/A will have an F distribution and we can then test whether or not there is an effect due to interviewers. We can estimate this effect (without assuming normality of the a_j's and ϵ_{ij}'s) by $\frac{B - A}{k'}$. Hochstim and Stock extend the model to deal with the case in which interviewers work in more than one block, the blocks being assumed to have been assigned at random to each interviewer. The block effects are also assumed to be additive and independent.

While in the examples given in their paper Stock and Hochstim claim that for the purposes of the analysis interviewer assignments can be assumed to be randomly distributed throughout the population, it is doubtful whether these requirements of random assignments will be met in most surveys. Usually interviewers are assigned certain compact segments of sampling units so that travel costs will be kept at a minimum. This is especially true in area sampling where sometimes widely separated clusters of sampling units are chosen. Some survey supervisors also feel that certain of their interviewing staff are more adept at obtaining interviews in certain segments of the population and assignments are therefore made in this decidedly non-random fashion.

It might also be said that the model used by Stock and Hochstim in their analysis of variance is not realistic inasmuch as it does not allow for interaction between interviewer and respondent. It is likely that a respondent would react differently to two different interviewers and a given interviewer will in turn be influenced in a different manner by different respondents.

Sukhatme and Seth (21, 1952), and Hansen and Hurwitz (22, 1951) have independently developed models, essentially equivalent, which take into account the interaction of interviewer with respondent. The following general mathematical model is given by Sukhatme and Seth:

$$y_{ijk} = x_i + a_j + \delta_{ij} + \epsilon_{ijk} \qquad \begin{cases} i = 1, 2, \ldots, l \\ j = 1, 2, \ldots, m \\ k = 1, 2, \ldots, n_{ij} \end{cases}$$

where x_i, a_j, and ϵ_{ijk} are defined as in the above model with $Ex_i = \mu$. δ_{ij} represents the interaction of the ith respondent with the jth enumerator. y_{ijk} can then be thought of as the value of the characteristic reported by the jth interviewer on the ith respondent for the kth occasion.

Assuming that the m interviewers to be used are fixed and that terms in 1/N are negligible in the case where there is a fixed population of N units the following expectation and variance can be derived:

$$E(\bar{y}_{\ldots}) = \mu + \sum_{j=1}^{m} \frac{n_{.j}a_j}{n}$$

$$V(\bar{y}_{\ldots}) = \sum_{i=1}^{l} \frac{n_{i.}^{2}}{n^2}\sigma^2 + \sum_{ij}\sum \frac{n_{ij}\sigma_{\delta_j}^{2}}{n^2} + \frac{\sigma_\epsilon^{2}}{n}$$

where $\bar{y}_{\ldots}$ is the total mean

$$\frac{1}{ml}\sum_{i=1}^{l}\sum_{j=1}^{m}\frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}} y_{ijk}$$

$$\sigma^2 = \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{N-1}$$ and $\sigma_{\delta_j}^{2}$ and σ_ϵ^{2} are similarly defined.

It is sometimes claimed that the effects of interviewers are self-cancelling and can therefore be ignored. This would obviously not be so if each $a_j = a$ (same common bias for all interviewers) but even if $\sum_j \frac{a_j n_{.j}}{n}$ were negligible for an adequately large group of interviewers, the sampling error,besides containing the intrinsic variation σ^2 will be inflated by the addition of terms in $\sigma_{\delta_j}^{2}$ and σ_ϵ^{2}. It will not be sufficient to know that the a_j's cancel each other but in each survey one should have some assurance that the contribution of response variation to the total variation is negligible.

If each unit is observed only once ($n_{ij} = 1$) it is not possible to get separate estimates of σ^2, σ_ϵ^2, $\sigma_{\delta_j}^2$ although we can get an estimate of σ_a^2 and upon making the appropriate normality assumptions we can test whether or not $\sigma_a^2 = 0$. The mean square between respondents within interviewers is now an estimate of $\sigma^2 + \bar{\sigma}_\delta^2 + \sigma_\epsilon^2$, where $\bar{\sigma}_\delta^2 = \sum_{j=1}^{m} \frac{\sigma_{\delta_j}^2}{m}$.

Sukhatme and Seth also show that in order to estimate separately the components σ^2, σ_a^2, $\sigma_{\delta_j}^2$, and σ_ϵ^2 we must have a situation where some of the units are reported on only once, some twice and those reported on twice being done either by the same interviewer or by two different interviewers.

It is obvious that in the usual surveys in which human populations are sampled this condition would be out of the question since it would be highly impractical to administer a questionnaire to the same respondent more than once. The best, therefore, that we can hope to do, if we assume Sukhatme and Seth's model to be the correct one, is to estimate σ_a^2 along with a linear combination of σ^2, $\sigma_{\delta_j}^2$, and σ_ϵ^2.

They also point out the fact that within any one stratum the test for differential interviewer effects will not have good discriminating power because of the small number of sampling units usually allotted to one stratum. However, if we pool the results for all strata we will no longer have a test of individual differences since we just get the average biases of the m enumerators averaged over all strata. Although this test may point up disagreement, to locate the disagreement it is still necessary to return to the individual stratum. On this basis plus the added travel costs necessary under these schemes Sukhatme and Seth disagree with Mahalanobis in the recommendation of the interpenetrating sample as an integral feature of a survey (see 23, 1948) but instead would relegate it to be used in pilot studies for improving the questionnaire or interviewing techniques. In its place they suggest that "adequate and effective supervision" of the field staff be instituted.

Hansen and Hurwitz (22, 1951) while using essentially the same mathematical model as Sukhatme and Seth use a different approach to the problem:

Let $\bar{y}$ = sample mean, which is used as an estimate of the true population mean $\bar{X}$,

$\bar{R}_y = E(\bar{y}) - \bar{X}$ is the response bias,

and the mean square error of $\bar{y}$ will equal $\bar{R}_y^2 + \sigma_{\bar{y}}^2$

where $\sigma_{\bar{y}}^2 = \dfrac{\sigma_y^2 - \sigma_{yI}}{n} + \dfrac{\sigma_{yI}}{k}$ for n respondents and k interviewers.

σ_{yI} is defined to be the covariance between responses obtained from different individuals by the same interviewer.

Hansen and Hurwitz's σ_{yI} corresponds to Sukhatme and Seth's σ_a^2 and $\sigma_y^2 - \sigma_{yI}$ to $\sigma^2 + \sigma_\delta^2 + \sigma_\epsilon^2$.

Hansen and Hurwitz now proceed, using a simple cost function, to determine the optimum number of interviewers for reducing the interviewer contribution, σ_{yI}, to the variance $\sigma_{\bar{y}}^2$. Both single and double sampling schemes are given for jointly reducing the bias component R_y^2 and the variance $\sigma_{\bar{y}}^2$. Also included in the paper is a discussion of the applicability of the specified mathematical model in which it is pointed out that what is needed is experimental evidence to determine whether or not the specified mathematical model is appropriate to approximate conditions found in any survey.

Suggested Methods and Their Limitations

It might at first appear feasible to make use of outside related information to *correct* for enumerator differences and the correction procedure might be stated as follows:

Suppose we know each interviewer's opinions with regard to a certain set of questions on a questionnaire to be administered. Suppose it is also assumed that there is a (linear) relationship between each enumerator's opinion and the responses of those that he interviews. Suppose the interviewers' responses can be assigned values $x_{ij} \begin{cases} i = 1 \dots k \\ j = 1 \dots m \end{cases}$ and corresponding to the jth question asked by the ith enumerator is y_{ij}. If we compute an analysis of covariance and from our analysis use the within interviewer regression coefficient b we can calculate adjusted means for the responses to each interviewer using $\bar{y}_i - b(x_i - \bar{X})$. In almost all cases, though, we are not interested in the result of each interviewer *per se* but in an average $\bar{y}$ summed over all interviewers. However, this $\bar{y}$ will be the same, whether or not any adjustment is made since $\Sigma b(x_i - \bar{x}) = 0$. If instead of using the within interviewer regression for adjustments we use the individual regressions b_i we will not get a zero average adjustment unless all the b_i's are equal but it will still be difficult to ascertain whether or not anything has been gained by making the adjustment since we are merely correcting to the mean of the group and there is no indication that this mean is any closer to the "true" value, than are the individual means. It is also very doubtful whether an interviewer's opinion and the response he elicits are linearly (or even curvilinearly) related. There is also no evidence to show that there is any bias due to the interviewer's opinions influencing responses. Any corrections, using the interviewer's opinions as the concomitant variable, can at best eliminate only a part of the interviewer effect; if they are used indiscriminantly to adjust to the mean value they give far worse results than if the responses had not been adjusted.

The assumption most likely to be violated in the models set out by both Hochstim and Stock and Sukhatme and Seth is the assumption that interviewer assignments are randomly chosen from the sampling units to be enumerated. As stated previously there is more likely to be some clustering of assignments within compact areas in order to save time and cut down travel costs. To require that each interviewer enumerate, or be available to enumerate, in all segments of the population would likely raise travel costs, so as to render the whole survey economically unfeasible. One of the main consequences, however, of the failure of the assumption of random assignments is that the interviewers are likely to be enumerating different segments of the population and interviewer differences now become confounded with segment differences. If it were possible to get an estimate as to what the segment effects were, they might be eliminated by means of a covariance analysis. Two methods of getting estimates of segment effects will be considered:

1. Use of an interviewer common to all segments whose results would be used as a measure of each segment's effect.

2. Use of outside information such as census data as a measure of each segment's effect.

It would probably be only feasible to use a "check" interviewer's results for some restricted group of interviewers, e.g., those interviewing in one city or one country or group of countries, since it would be required that the "check" interviewer take a certain number of interviews in the same segments as each other interviewer. Unless the scheme were restricted to a small number of segments it would be physically impossible for the "check" interviewer to complete his assignments in the time allotted for the survey. One objection to this scheme is that with the present size of segment being used it is doubtful whether the "check" interviewer would be able to take enough interviews in any one segment to provide a usable estimate of its effects. If it were possible to overcome this objection one would use "between interviewers" error of estimate as the greater mean square in the F ratio and the "between respondents within interviewers" error of estimate as the lesser mean square. One further assumption, besides the usual ones necessary for using the F ratio, is that there is no interaction between the "check" interviewer and segments. If an interaction were present we would not only be getting a measure of the effect of the segment but also a component due to the variable effect of segments upon the "check" interviewer.

The use of outside information such as census tract or block statistics data provided by the Bureau of the Census is a much more hazardous type of adjustment since we are tacitly assuming that there is a simple relationship between the characteristics one is studying and the information at hand. Even if this assumption were warranted the adjustment by covariance would be extremely tedious and time-consuming since multiple and perhaps curvilinear relationships would likely be assumed, depending upon the characteristic being measured. It is doubtful whether we could assume the segments to be homogeneous in all characteristics merely because we had adjusted for differences in certain other characteristics. More efficient use can probably be made of this type of data if it is used as a check to be compared with interviewers' results when enumerating the same characteristics. Care should be taken when using outside information in this regard since it often becomes obsolete within a short period of time after publication.

One method of assigning interviewers at random to the sampling units so that interviewer comparisons can be made but at the same time grouping them in relatively compact segments would be to use a "partially balanced incomplete block design" to lay out interviewer assignments. This would require that segments be large enough to afford two interviewers one or more days work. It is also required that each interviewer be available to interview in any segment of the population and for this reason such a plan would probably be restricted for use in comparing interviewer results within a large city or a county. It is not necessary that all interviewers interview the same number of respondents but it is necessary that they interview in the same number of segments, this number being specified by the design.

A partially balanced incomplete block has the following properties (see 24, 1952) :

1. There are t treatments in b blocks of k plots, and r replicates of

each treatment.

2. With reference to any specified treatment, the remaining (t-1) fall into m sets, the ith of which occurs with the specified treatment in λ_i blocks and contains n_i treatments, the number n_i being the same, regardless of the treatment specified.

3. If we call the treatments that lie in a block λ_i times with a specified treatment θ , the ith associates of θ, the number of treatments common to the ith associates of θ and the jth associates of ϕ, where θ and ϕ are kth associates is p^k_{ij} , this number being the same for any pair of kth associates.

In the context of this paper we would designate t interviewers in b segments consisting of k (= 2) interviewers each, with each interviewer interviewing in r segments.

For example, to compare 9 interviewers in blocks of 2 our design would have the following parameters:

$t = 9$, $b = 18$, $k = 2$, $r = 4$, $n_1 = 4$, $\lambda_1 = 1$, $n_2 = 4$, $\lambda_2 = 0$,

$p^1_{ij} = \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix}$, $p^2_{ij} = \begin{pmatrix} 2 & 2 \\ 2 & 1 \end{pmatrix}$.

This particular design would require that there are at least 18 segments or blocks and that each interviewer interview in 4 segments.

An example is now given of how a design with these particular parameters would be applied in a practical situation. We will suppose that segments have been chosen and that sampling units within each segment specified in advance. For our design to be feasible in practice it would be required that each segment be large enough to provide two days interviewing for each of the two interviewers working within a segment. This will allow each interviewer to attempt to interview on the second day those respondents missed in the first attempt.

The nine interviewers are first assigned randomly into pairs as specified by the design. Suppose the pairing is as follows:

(A, B), (B, H), (E, F), (B, E), (D, G), (H, I), (A, C), (C, F), (E, H), (B, C), (D, F), (G, I), (A, D), (C, I), (F, I), (A, G), (D, E), (G, H).

Note that we have 18 interviewer pairs with each interviewer appearing in 4 different pairs. If there are more than 18 segments to be sampled we may choose 18 of these at random.

Each of the 18 pairs of interviewers is now assigned at random to the segments. We might end up with the following distribution of pairs:

Segment	1	2	3	4	5	6	7	8	9
Pair	AB	AC	AD	AG	BC	BE	BH	GF	CI

Segment	10	11	12	13	14	15	16	17	18
Pair	DE	DF	DG	EF	EH	FI	GH	GI	HI

Within each segment serial numbers are assigned to the sampling units in a systematic fashion. One interviewer of a pair can then be assigned at random to the even numbered units within the segment and the other to the odd numbered units.

Since the experimental unit is the sum of all sampling units for a given interviewer in a given segment it is not necessary to assign the interviewer to each of the sampling units at random but merely to one of the two experimental units in each segment at random. Each interviewer's assignment is thus completely determined and within these specifications the interviewing procedure can be carried out as usual. The shortest routes within and between segments can be determined and the interviewing can be scheduled so as to minimize travel time. It is not necessary that both interviewers be in a segment at the same time.

To illustrate how the analysis would be carried out suppose that it is desired to test whether or not there is any difference in the percentage of "don't know" (D.K.) responses the nine interviewers received to a certain question. In this case some action might be taken if differences are detected since those interviewers receiving a high number of D.K.'s might not be using probes when they should be. On the other hand interviewers receiving an unduly low number of D.K.'s may be forcing the respondent's replies into categories in which they don't really belong.

The example has been constructed so that one interviewer, F, reports an inordinately high percentage of D.K.'s while interviewer D shows a very low percentage. Although a transformation should be applied to the percentages before computing an analysis of variance this won't be carried out here since the data is used merely to illustrate a technique.

The model used for the analysis of intra-block estimates will be the simple additive model

$$y_{ij} = \mu + b_i + t_j + \epsilon_{ij}$$

When there is no appreciable decrease in experimental error brought about by the grouping of units into incomplete blocks it is desirable to make use of inter-block estimates in forming a combined estimate of treatment (interviewer) effects. To make use of this estimate it will be necessary to assume that segments are a random sample from some population of segments. In our case this is not a severe restriction since this assumption is generally satisfied by the survey design.

The data are given below:

1	2	3	4	5	6
A 8	C 5	A 4	G 9	B 7	E 5
B 4	A 7	D 0	A 10	C 7	B 7

7	8	9	10	11	12
H 9	C 6	I 8	DD 0	F 13	D 3
B 10	F 16	C 6	E 2	D 3	G 7

13	14	15	16	17	18
E 6	H 9	F 15	G 6	I 7	I 9
F 17	E 5	I 5	H 13	G 4	H 10

The analysis of variance will take the following form:

Source	df	ms
Segments (ignoring interviewers)	17 (b-1)	
Interviewers (eliminating segments)	8 (t-1)	34.5**
Error	10 (bk-b-t+1)	.456
Total	35 bk - 1	

The intra-block estimates of interviewer effects are:

A	B	C	D	E	F	G	H	I
9.2	6.5	4.3	1.2	3.2	20.2	4.3	9.9	6.7

The efficiency factor of a design of this type is 50% of what it would have been if interviewers had been compared in complete blocks. Since it is desirable to make use of inter-block information the following analysis of variance will be necessary for estimating weighting factors:

Source	df	ss	ms
Segments (eliminating interviewers)	17	93.441	5.495
Interviewers (ignoring segments)	8	463.22	
Error	10	4.559	.4559

The combined estimates using inter-block information will be found to be

A	B	C	D	E	F	G	H	I
7.9	6.7	5.9	3.1	4.7	15.6	5.2	9.0	7.4

The methods of computation may be found in (24, 1952). The computations were not shown here since the example is a straightforward application of the analysis for a partially balanced design.

If discrepancies in results are not so obvious as in this example one might make use of Tukey's method (25, 1949), using as an approximate estimate of error variance:

$$\frac{\text{error mean square}}{\text{bk}} .$$

One is then able to test whether or not the most discrepant means differ significantly from the group mean and to group the means into homogeneous classes or groups.

What action to take when the interviewer means fall into two or more equal groups will differ in different situations. One would probably examine each group to look for such common factors as training, supervision, or personality. On the basis of this examination it may be possible to come to some conclusion as to which group's results are discrepant and what method of correction to use.

It is doubtful whether the conditions under which any detection method might be used will be completely met in all surveys. The main difficulty faced is the fact that interviewer assignments are not randomized and that any attempt to randomize them will result in a rise in travel costs. Since for most organizarions travel costs form a very large proportion of interviewing costs there is an understandable reluctance to incorporate a method into a survey which will raise costs unless some benefits are assured from the method. If a survey organization is able to ascertain how discrepancies in results arise and feel that the information arising from a detection scheme will be of some use in re-training their interviewing staff for future surveys then such a scheme may be of value to them. As pointed out before, such an analysis will not tell them whether or not the interviewing staff as a whole is getting biased results because, for example, of a faulty training method, unless other information is available for comparison with the results.

What is probably required before detection methods for interviewer differences can be utilized more fully and efficiently is empirical verification of the assumed mathematical models along with more fundamental research into the sources and ramifications of interviewer bias.

Bibliography

1. Deming, W. E.
On errors in surveys.
Am. Soc. Rev. 9:359-369. 1944.

2. Hyman, H.
The isolation, measurement, and control of interviewer effect.
Soc. Sci. Res. Coun. Items 3, 15 Vol. 3, No. 2. 1948

3. Calahan, D., Tamulonis, V., and Verner, H. W.
Interviewer bias involved in certain types of opinion survey questions.
Internat'l J. Opinion and Attitude Res. 2:63-77. 1947.

4. Manheimer, D., and Hyman, H.
Interviewer performance in area sampling.
Public Opinion Quarterly 13:83-93. 1949.

5. Sheatsley, P. B.
The influence of sub-questions on interviewer performance.
Public Opinion Quarterly 13:310-313. 1949.

6. Stember, H., and Hyman, H.
Interviewer effects in the classification of response.
Public Opinion Quarterly 13:669-679. 1950.

7. Smith, H. L., and Hyman, H.
The biasing effect of interviewer expectations on survey results.
Public Opinion Quarterly 14:491-506. 1950.

8. Feldman, J. J., Hyman, H., and Hart, C. W.
A field study of interviewer effects on the quality of survey data.
Public Opinion Quarterly 15:734-762. 1951.

9. Yates, F., and Watson, D. J.
Observer's bias in sampling-observations on wheat.
Empire J. Exp. Agr. 2:174-177. 1934.

10. Yates, F.
Some examples of biased sampling.
Annals of Eugenics 6:202-213. 1935.

11. Cochran, W. G., and Watson, D. J.
An experiment on observer's bias in the selection of shoot-heights.
Empire J. Exp. Agr. 4:69-76. 1936.

12. Cochran W. G.
The estimation of the yields of cereal experiments by sampling for the ratio of grain to total produce.
J. Agr. Sci. 30:262-275. 1940.

13. King, A. J., and Jebe, E. H.
An experiment in pre-harvest sampling of wheat fields.
Iowa Agr. Exp. Res. Bul. 273:624-649. 1940.

14. King, A. J., McCarty, D. E., and McPeek, M.
An objective method of sampling wheat fields to estimate production and quality of wheat.
U. S. Dept. Agr. Tech. Bul. 814:1-87. 1942.

15. Houseman, E. E., Weber, C. R., and Federer, W. T.
Pre-harvest sampling of soybeans for yield and quality.
Iowa Agr. Exp. Sta. Res. Bul. 341:808-826. 1946.

16. Homeyer, P. G., and Black, C. A.
Sampling replicated field experiments on oats for yield determinations.
Soil Sci. Proc. 11:341-344. 1946.

17. Sukhatme, P. V.
The problem of plot size in large-scale yield surveys.
Am. Stat. Assoc. J. 42:297-310. 1947.

18. Durbin, J. J., and Stuart, A.
Differences in response rates of experienced and inexperienced interviewers. (with discussion)
Royal Stat. Soc. J. 114: Series A, Part II:163-205. 1951.

19. Mahalanobis, P. G.
Recent experiments in statistical sampling in the Indian Statistical Institute. (with discussion)
Royal Stat. Soc. J. 109, Series A:325-370. 1946.

20. Stock, J. R., and Hochstim, J. A.
A method of measuring interviewer variability.
Public Opinion Quarterly 15:322-334. 1951.

21. Sukhatme, P. V., and Seth, G. R.
Non-sampling errors in surveys.
Indian Soc. Agr. Stat. J. 4:5-41. 1952.

22. Hanson, M. H., Hurwitz, W. N., Marks, E. S., and Maudlin, W. P.
Response errors in surveys.
Am. Stat. Assoc. J. 46:147-160. 1951.

23. U. N. Sub-commission on Sampling.
Recommendations on the preparation of reports of sample surveys. 1948.

24. Kempthorne, O.
The design and analysis of experiments.
Chapter 27. John Wiley and Sons. 1952.

25. Tukey, J. W.
Comparing individual means in the analysis of variance.
Biometrics 5:232-242. 1949.

26. Kempthorne, O.
A class of experimental designs using blocks of two plots.
Annals Math. Stat. 24:76-84. 1953.