ON SOME ASPECTS OF RESPONSE ERROR MEASUREMENT[1,2]

William G. Madow, Stanford Research Institute

1. Introduction. Both response error and response bias[3] occur in almost every survey, whether a survey of a sample or a survey of an entire population.

The understanding and measurement of response error are essential for four main purposes:

(1) To improve the assessments of the accuracy of data provided by a particular survey, and to determine how much weight can be given those data in procedures for making conclusions, decisions, or actions in which the data are inputs,

(2) To determine how to improve the making of the survey and the estimates based on it,

(3) To improve the allocation of resources to the different parts of a survey, viewed as an economic enterprise, in order to maximize the information provided by a survey for a given cost or to minimize the risks of conclusions, decisions, or actions based in whole or part on the survey, and

(4) To contribute to the accumulation of information on response error for use in other surveys and in research on survey making.

In many surveys response errors are large; usually, they are measured or reduced only with considerable effort and cost. In some surveys, the main component of response error is response variance, which can be reduced by technical improvements in the size or method of making the survey. In other surveys, the response bias is large. To reduce response bias is often far more costly than to reduce response variance; sometimes the response cannot be sufficiently reduced to regard it as small. Sometimes, the knowledge and understanding of the bias are adequate for taking it into account when conclusions, decisions or actions are taken; sometimes other evidence than the survey itself will keep the response bias from leading to error; sometimes, in time series, the changes in the data are sufficiently greater than the changes in the biases for estimates of change to be relatively unaffected by response bias. To know whether the response bias will produce errors in analyses requires knowing something of the size and nature of the response bias.

To measure response variance and response bias, to determine how they depend on various parts of the survey, to determine how they are related to respondent and interviewer characteristics, to determine how they are related to other variables on which information is obtained in the surveys, to determine how they change over time, to determine how and at what cost they can be reduced, and to determine how much they increase the risks of conclusions, decisions and actions utilizing the survey; all these seem essential to the improvement of survey making and use.

In this paper we emphasize:

(1) The possibility of using double sampling methods for reducing response bias,

(2) The contribution of explanatory variables to the understanding of response error, and to the reduction of response error,

(3) The importance of studying as behavioral science models the choice mechanism used by the respondent in selecting one of the possible replies that he can give when a question is put to him, not only in order to understand response error but to reduce it, and to lower the cost of double sampling methods for reducing response error.

We also discuss the concept of the exploratory survey, a type of survey that we distinguish from the analytic, general purpose or special purpose surveys. Since the treatment of response error may be quite different in surveys for time series or in panel surveys than in one-time surveys, we also briefly consider that classification.

A suggestion that we make, although we do not implement, is an attempt to resolve an issue that often arises between the survey designer and the survey user. This issue is the contradiction between the user's knowing he cannot state precisely how the data will be used, and his consequent wish that he be provided with generally good data; and the designers knowing he cannot provide data adequate for all uses and his consequent wish that he be provided with exact statements of how accurate specified data must be for the uses to be made of them.

The suggestion is that the designer and user, instead of discussing accuracy begin by considering some perhaps oversimplified models of using the data, discuss the requirements on the data resulting from the desire to achieve certain risk levels through the use of these models, and conclude by adjusting the accuracy levels to account for other possible uses not so easily specifiable.

[1] This paper owes much to the work, published and unpublished, of the staff of the Bureau of the Census. Of the many papers they have published in this area only one is mentioned: Measurement Errors in Censuses and Surveys by Morris H. Hansen, William N. Hurwitz and Max A. Bershad, Bull. Int. Stat. Inst. Vol. 38 (1961) pp. 359-374.

[2] This research was supported by National Science Foundation Grant GS-83.

[3] We shall refer to response variance and response bias collectively as response error.

2. Classifying surveys by their uses.

We first discuss two classifications of surveys that are related to the uses of surveys. The word "surveys" as used here includes not only sample surveys, but also surveys of an entire population or universe.

One classification of surveys is by whether the surveys are "one-time" surveys, or whether the surveys are made to provide data for a time series, and in the latter case, whether the survey is a panel survey made at a fixed number of time periods with complete or partial renumeration of the members of the sample at the time periods.

Another classification of surveys is by whether the survey is intended to provide data for general use, such as data provided by the Bureau of the Census or the Bureau of Labor Statistics or various other government agencies, or whether data are to be provided for special purposes, such as the data provided by marketing surveys made for a specific company in order for it to decide what action it wishes to take with respect to its products, or whether the survey is an analytical survey made in order to test various specified social science hypotheses such as might occur in studies intended to test hypotheses concerning the nature of voting behavior or whether the survey is an exploratory survey, i.e. a survey made in order to obtain information concerning which variables are important in characterizing one or another kind of behavior.

The classifications of surveys made above are not classifications into mutually exclusive classes. Many surveys partake of several of these aspects. From the point of view of survey design however, these classifications lead to different designs, and consequently the consideration of these classifications is relevant to a discussion of survey design.

Let us discuss the differences among these various types of surveys in relation to a common body of information. For this purpose the topography of a geographical area provides a useful illustration.

In a general purpose survey, the topographical features such as mountains, rivers, etc. would have been identified; for each of these features specific items such as heights, lengths, etc. would be measured by some procedure. The list of topographical features and the measurements together with indications of how accurate these measurements would be would then be published as a general purpose set of data available for any user who could himself decide the extent to which he wished to rely or make use of these data. In making the survey, certain users and their needs might have been considered but, in addition, the data be for general use. The data may make it possible for some decisions to be made but might not be adequate for others - perhaps because of the wide focus rather than a narrow focus.

In a special purpose survey, attention would usually be put on one or more of the topographical features; the objective might be to learn whether one mountain is larger than another without it being terribly important to know the exact size of either of the two mountains, or it might be necessary to learn something of the slope of the mountain that would go beyond the detail published in a general purpose study, perhaps, in order to estimate the cost of moving material from one level to another of the mountain. A frequent characteristic of the special purpose survey is that it is intended to provide data required to make a decision or take some action.

An analytical survey might be concerned with testing hypotheses concerning the ages of the mountains, or a hypothesis concerning the nature of the phenomena that led to the existence of some of the existing topographical features, and the prediction of what these topographical characteristics might be at some future time period because of the nature of the fundamental processes creating change in the topographical characteristics.

The exploratory survey would be concerned with attempting to determine what were the real topographical features themselves, which of these features best characterized the area in question, and develop hypotheses concerning the underlying processes.

The main reason for discussing this classification at this point is that different kinds of samples and different types of analysis would be appropriate for the different types of surveys identified in the classification. For example, a special purpose survey with a very specific objective can often have far smaller size and be less rigorously made than a general purpose survey of the same area intended to provide information suitable for many uses, uses which may not have even been specified at the time that the survey was made. While this does not mean that all general purpose surveys must be of exceptionally high quality, it does imply that if a survey can be done for specific purposes the costs can be reduced.

The distinction between analytical surveys and either general or special purpose surveys is not a distinction between whether one wishes to estimate relations or only to estimate specific numbers. Both general and special purpose surveys may well have as their major objectives the estimation of both numbers and relations, or even only relations. Similarly general purpose studies, special purpose studies and analytical studies all may be intended to provide data to serve as a basis for decisions. The main difference between analytic surveys and either general or special purpose surveys is that analytical surveys are concerned with the processes that led to the data that are obtained or to the relationships that are obtained, whereas special purpose and general purpose surveys are primarily concerned with the data or relationships themselves and their uses in decision and action.

It is almost inevitable that analytical surveys would contain what are sometimes called "soft" data as well as the relatively "hard" data that are often obtained in special or general purpose surveys, although this is not meant to imply that only hard data are obtained in such surveys.

From a design point of view, the main difference between exploratory surveys and either general or special or analytical surveys is that in the case of exploratory surveys it is so much more important to sample all the possible sources of variability of the data to be obtained whereas in any of the other three the sources of variability can be limited by setting forth some more restrictive objectives that can be stated for an exploratory survey.

3. Steps in survey making related to response error.

The steps of the survey that include the selection of a respondent, the design of a questionnaire, the selection of an interviewer, and the bringing of these three entities together may be viewed as the planning of a choice experiment in which the respondent presented with a combined stimulus of the question or questionnaire and interviewer in the environment in which the interview takes place chooses a response either from a pre-assigned list of possible responses or from a set of alternatives among which the respondent decides to make the choice.

Not all elements of this experiment need be present. For example, it may be that instead of the respondent there is a file of information concerning people or businesses, and the interviewer, who in this case is the person who records the information, may abstract the information from the record onto the form. This procedure need not be error free. For example, if the forms from which the information is being extracted are the notes of physicians in medical records the abstraction of the information will be subject to many sources of error.

The interview may be conducted by mail or by telephone. In the former case the respondent may have an opportunity to read the entire questionnaire before he responds to individual questions, whereas either in a personal or telephone interview he is asked the questions sequentially and his responses may be affected by that fact. There may be no written questionnaire; the interview may be unstructured, and may be essentially non-directive as in the case of psychiatric interviews or similar types of interviews that occur in surveys.

The structure of the interviewing situation is that for each question, the respondent makes a choice from a more or less specified set of alternatives. The choices, (responses) are usually made sequentially although, as in the case of a mail survey, sometimes they are made collectively. The set of alternatives may be spelled out as in the case of a forced choice question, or to some extent defined by the respondent himself - usually without stating the set to the respondent.

The information provided by the respondent may be categorical as for example in the case of sex or it might be more nearly continuous as in the case of age or income. No distinction is being made according to the type of information provided by the respondent.

4. The accuracy required of estimates. Small analytic models.

The requirements of the analysis of the information provided by the survey must indicate the accuracy demanded of that information. Yet, statisticians and survey designers have long had the experience that it is difficult, if not impossible to expect the users of data to indicate the uses sufficiently precisely for the accuracy requirements on the survey to be possible to determine. Most frequently, such accuracy requirements have been arrived at by a process of negotiation in which the user will first suggest it would be nice if the data were completely accurate, the designer will inform the user of the cost of such an achievement, even assuming it to be possible which usually it is not, the user will then reduce the accuracy requirements he would like to place on the data, the designer will indicate the costs and feasibility of the reduced level of accuracy, etc. until either an agreement is reached or the decision is made either not to obtain the data or to obtain it by other means. One reason why it is so often difficult to state the required accuracy of the data is that the uses are themselves not too specifiable especially at the time of designing the survey. For example if one is talking about television ratings then it is well known that many decisions are made just by looking at the television ratings, but that many other aspects of information are used than those provided by the survey itself. Similarly if a government policy is to be based on the results of a survey, that policy will take into account far more then the numerical results of the survey; political aspects and fundamental economic considerations may also play a very large part. These considerations are even more important when one considers analytical and exploratory surveys in which the nature of the analysis will be affected by the results obtained in the survey itself. Between the date on which the design is fixed and the date on which the data begin to be available, several months may pass, and the needs may evolve and change. Furthermore, as the analysis evolves the needs and accuracy requirements may both change.

Even if the needs do not change, it often occurs that when decisions on the accuracy required are obtained by the process of negotiation between the survey user and the survey designer, it turns out that the negotiations either did not cover all the actual uses to be made of the data, or that the user did not fully understand to what he was agreeing.

Very often, once data have been collected then despite all the limitations placed upon the use of

the data by the survey designer or statistician, the user often feels that since the data are the best he has available, he must use them and rely primarily on the consistency of the conclusions obtained from the data with other sources of information, or the internal consistency of the conclusions as the basic support for his analysis. It is true that if the data are highly variable then there will tend to be a sufficiently large number of inconsistencies for the user to become concerned and to limit the conclusions that he was prepared to draw from the data. But, when the data are biased, there may be few inconsistencies, and yet the survey as a whole is wide of the mark. A sufficiently sophisticated user who makes comparisons with relatively unbiased outside sources of data, and evaluates these data themselves will often use the survey in a useful way. However, it must be emphasized that it is the user whose judgment becomes critical rather than the objective properties of the survey data themselves.

Sometimes when a time series is being analysed it is felt that biased data may be used because the effects of the bias will disappear or be greatly reduced when investigating the changes over time. The biases may be due to response, as in income estimates, or to the use of a study population that differs from the population concerning which conclusions are to be drawn, as in the case of TV ratings. But often the biases will change over time, and may even increase in relative importance. Just as we attempt to measure the accuracy of estimates at a given time so must such estimates be attempted over time; in particular the changes in response bias should be estimated even if it is believed that response variance is stable.

There is an intermediate level that might help determine the required accuracy of the data without committing the user to use methods of analysis that are too simple or inappropriate for his purposes; this method is the construction of small models. These models may be analysed mathematically or by simulation. But they will lead to conclusions on the accuracy required of the data so far as their requirements are concerned, and such analyses will often provide an adequate basis for deciding the accuracy to be required of the data.

5. Questionnaire and Interviewer Effects on Response Error.

We have suggested how it may be possible more often than now is done, to arrive at conclusions concerning the accuracy of the estimates desired from a survey. Let us now turn to the ways in which accuracy may be measured and to the factors on which the accuracy depends.

One of the most important factors to consider, and one of the most difficult properly to assess, is the choice of questionnaire. Most discussions of questionnaire construction come down to telling the questionnaire constructer to be wise and to be careful, and point out that unless one is wise and careful variance and biases will result. Experience indicates however, that despite all the wisdom and judgment of the past, every new questionnaire presents a new challenge, and the ability of respondents to find weaknesses in the questionnaire continues to be effective even after serious and able attempts of questionnaire constructors to find and remove biases before the respondents show him the weaknesses. Pretests or pilot studies of questionnaires are known to be indispensable, and often these pretests and pilot studies include alternative questionnaires. Improvements of questionnaires are usually costly to demonstrate; and often it is difficult to justify the improvement in information in terms of the extra cost and time and, perhaps, the more competent interviewers and greater training that are required. Even when two or more versions of a questionnaire are used in the same study so that differences primarily due to the questionnaires can be isolated, and the versions are distributed among the sample of respondents so that valid comparisons can be made, the versions may have the same essential biases. Biases because of unwillingness to report, for example, may not be reduced sufficiently by an improvement in a questionnaire; similarly inability to report because of forgetfulness may be only partially reduced by the use of questions intended to stimulate the memory; the different versions of the questionnaire may yield data more similar to each other than any is to the true data that one would like to elicit.

It is rare that one-time surveys can adequately develop questionnaires, especially since so few tests of questionnaires are conducted in a way that provides objective evidence.

Once the questionnaire or questionnaires have been selected, the issue becomes that of the accuracy of the information requested on the questionnaires.

Other important topics in survey making that require further research in spite of the large and costly efforts that have been made are the training of interviewers and the manner of conducting interviews. Interviewer training has been difficult to evaluate because of the very high cost of experimental studies that are generalizable. Whether the interviewer should ask questions exactly as they appear on the questionnaire or be allowed to vary from the questionnaire seems to depend on the survey and the organization conducting the survey. It is well known that even slight deviation from questions as worded may produce unexpected effects in the responses obtained. Yet, whatever may be the emphasis placed on the importance of the interviewer not adjusting the questionnaire, changes are often made by the interviewer even when this goes counter to the instructions. Even when no change is explicitly made, the interviewer's attitude toward the question communicates itself sufficiently to the respondent to alter the meaning of the question. For example, the interviewer may feel that questions on income should not be asked, and have difficulty in asking such questions; even when asking the

questions the interviewer may ask them sufficiently tentatively or concernedly so that the respondent becomes aware without any words being spoken that make the interviewer's attitude explicit. Similarly an interviewer with long experience in using a particular questionnaire may stop asking all the questions or using all the aides provided because of sufficient familiarity with the questionnaire to believe that these practices are unnecessary. For example, the interviewer may have been given a list to show to the respondent at a certain point but instead of showing the list to the respondent, the interviewer may eventually memorize the list and repeat it - perhaps incorrectly.

Similarly the problems that the interviewer faces in attempting to communicate with the respondent could be the subject of further study. Many respondents will not refuse, but will lower the quality of information they provide by the way in which they participate. The interviewer faced with a resentful or hurried or troubled respondent will often be affected by the respondent and the quality of information will suffer thereby. Similarly, the interviewer, in obtaining information from a proxy respondent, will find that it is more difficult to ask probing questions concerning the desired respondent in any fruitful fashion.

Under these conditions, the measurement of response variance and response bias become almost indispensable in a survey aimed at more than the establishment of very large differences.

6. Response Variance, Response Bias and Response Error.

It will be helpful at this point to introduce some definitions and symbols in order to make precise the analysis with which we will be concerned.

Suppose that we are dealing with the population of N elements, denoted by $1,2,\ldots,N$.

Let μ_i, $i=1,\ldots N$ be the "true value" of a variable for the ith element of the population.

Let x_i, $i=1,\ldots N$ be the random variable that is the choice of the respondent i, if respondent i is asked the question for which the true value for that respondent is μ_i.

Let a_i, $i=1,\ldots N$ be the expected value of x_i, i.e., $E\, x_i = a_i$.

Then, for the ith respondent, the variance of response, RV_i, is the expected value of the square of the difference $x_i - a_i$, i.e.

$$RV_i = E\,(x_i - a_i)^2$$

and the response bias, RB_i, is the difference between the expected response and the true value i.e.

$$RB_i = a_i - \mu_i$$

Finally, the mean square error of response, M_i, is the expected value of the square of the difference $x_i - \mu_i$, i.e.

$$M_i = E(x_i - \mu_i)^2 = E(x_i - a_i)^2 + (a_i - \mu_i)^2.$$

We shall not assume that the errors of response are uncorrelated. Also, we shall be assuming the existence of a "true value" even though, in practice a "true value" may not exist. For example, in cases of illness, the differences among physicians on whether a person is ill often are sufficiently great to cause us to remember that illness is itself a continuous variable and that what illness is will often be a matter of opinion. To define "illness" so that except for measurement errors among doctors, the same diagnoses would be reached would be most difficult. This problem is serious enough when it comes to ordinary medical conditions but in dealing with psychiatric problems the differences are much greater. Similarly when one is dealing with opinion and attitude questions the true value may well be a true probability distribution of opinions or attitudes of the person. Even in deciding whether a person is unemployed, or is a member of the labor force, large elements of opinion are present; whether there is really a true value for these variables is often doubtful. Sometimes, as in the case of unemployment, there may be knowable true values for many elements of a population without there being true values for all the elements of a population. Many are certainly employed, many are certainly unemployed, many are certainly not in the labor force, but many also are in the fringe groups where the true values are uncertain. To change definitions to eliminate fringe groups is not necessarily to increase the information provided by the survey.

Let us suppose now that the objective of a survey is to estimate a function,

$$f = f\,(\mu_1, \mu_2, \ldots, \mu_N)\,,$$

of the "true values", $\mu_1, \mu_2, \ldots, \mu_N$.

For this purpose, an estimator, g', based on sample values $x'_1, x'_2, \ldots, x'_n$, i.e.

$$g' = g(x'_1, x'_2, \ldots, x'_n)\,,$$

is defined.

Let $E\,(g' \mid \mathcal{S})$

be the expected value of g', given the selected sample. Then, the response variance of g', $RV_{g'}$, is defined to be

$$RV_{g'} = E\sigma^2_{g'|\mathcal{S}} = E\left\{ E\left[(g' - E(g'|\mathcal{S}))^2 \,\middle|\, \mathcal{S} \right] \right\}$$

The response bias, $RB_{g'}$, is by definition

$$RB_{g'} = Eg' - f$$

When $E(g'|\mathcal{S})$ is evaluated it will be some function $h'_a = h(a'_1,\ldots,a'_n)$, i.e.

$$E(g'|\mathcal{S}) = h(a'_1,\ldots,a'_n) = h'_a$$

Define h'_μ to be the same function of $\mu'_1, \mu'_2,\ldots, \mu'_n$ that h'_a is of $a'_1,\ldots, a'_n$.

Often, $h(\mu'_1,\mu'_2,\ldots,\mu'_n)$ will be the estimate of f that would have been used, had there been no response error. Sometimes, another function would have been used, possibly $g'_\mu = g(\mu'_1, \mu'_2,\ldots, \mu'_n)$. Let us denote by $k'=k(\mu'_1, \mu'_2,\ldots, \mu'_n)$ the function that would be used to estimate f if there were no response error.

Then,

$$* \quad E\left[E(g'|\mathcal{S}) - Ek'\right]^2 = E\left[E(g'|\mathcal{S}) - k'\right]^2 + 2E\left[E(g'|\mathcal{S}) - k'\right]\left[k' - Ek'\right] + E(k'-Ek')^2$$

and we shall define the expected square sample bias, $SB^2_{g',k'}$ by the equation

$$SB^2_{g',k'} = E\left[E(g'|\mathcal{S}) - k'\right]^2,$$

the regression coefficient of $E(g'|\mathcal{S}) - k'$ on k', $\beta_{E(g'|\mathcal{S}) - k',\, k'}$, by

$$\beta = \beta_{E(g'|\mathcal{S}) - k',\, k'} = \frac{E\left[E(g'|\mathcal{S})\, k'\right]\left[k' - Ek'\right]}{E(k' - Ek')^2}$$

and the mean squared error had there been no response error, $MS_{k',\,f}$, by

$$MS_{k',\,f} = E(k' - f)^2 = E(k' - Ek')^2 + (Ek' - f)^2$$

As mentioned above k' will often be either g'_μ or h'_μ.

Let $SB^2_{g',k'}$ be partitioned into

$$** \quad SB^2_{g',\,k'} = E\left[E(g'|\mathcal{S}) - k') - (Eg' - Ek') - \beta(k' - Ek')\right]^2 + (E\,g' - E\,k')^2 + \beta^2 E(k'-Ek')^2$$

$$= S^2_{E(g'|\mathcal{S}) - k',k'} + (E\,g' - E\,k')^2 + \beta^2 E(k'-Ek')^2$$

Since

$$E(g' - f)^2 = \sigma^2_{E(g'|\mathcal{S})} + E\,\sigma^2_{g',\mathcal{S}} + (Eg' - f)^2$$

$$= E\,\sigma^2_{g',\mathcal{S}} + E\left[E(g'|\mathcal{S}) - f\right]^2$$

and since

$$E\left[E(g'|\mathcal{S}) - f\right]^2 = E\left[E(g'|\mathcal{S}) - E(k')\right]^2 + 2E\left[E(g'|\mathcal{S}) - Ek'\right]\left[E\,k' - f\right] + (E\,k' - f)^2$$

it follows that from * and ** that

$$E\left[E(g'|\mathcal{S}) - f\right]^2 = S^2_{E(g'|\mathcal{S}) - k',k'} + (\beta+1)^2\sigma^2_{k'} + (Eg' - f)^2$$

Hence we arrive at the fundamental formula for the mean square error of the estimator g' about the value estimated, f:

$$E(g'-f)^2 = E\sigma^2_{g'|\mathcal{S}} + S^2_{E(g'|\mathcal{S}) - k',k'} + (\beta+1)^2\sigma^2_{k'} + (Eg' - f)^2$$

Of the four terms, the first is the response variance and the last is the square of the bias. We have chosen to express the two middle terms using the regression of $E(g'|\mathcal{S}) - k'$ on k' since the within sample bias will so often be a function of k'.

It is easy to express these two middle terms using the regression of $E(g'|\mathcal{S})$ on k'. It will be noted that, especially if k' is an unbiased estimate of f, the occurrence of a negative regression coefficient of $E(g'|\mathcal{S}) - k'$ on k' can yield values of $E(g'-f)^2$ that are smaller than $E(k' - f)^2$.

The dominant term in $E(g' - f)^2$ may well be $(Eg' - f)^2$, the overall bias, squared. For this reason we turn in Section 7 to the use of double sampling methods of eliminating or reducing the bias.

7. The reduction or elimination of response bias by double sampling.

In order to avoid complications unnecessary for the discussion of the ideas we shall deal in this section with the simplest case, that in which a simple random sample has been selected from which a simple random subsample is selected for which estimates of the bias are obtained, and a difference estimate is used.

The true values for the population are $\mu_1, \mu_2,\ldots, \mu_N$. The objective of the survey is to estimate $\bar{\mu}$, the arithmetic mean of $\mu_1,\ldots, \mu_N$.

The responses of the elements of the sample are designated by $x'_1, x'_2,\ldots, x'_n$ and

$$E(x'_i \mid i) = a'_i,$$

where a'_i is also a random variable because of the use of simple random sampling.

Let

$$\bar{x}' = \frac{x'_1 + \ldots + x'_n}{n},$$

$$\bar{a}' = \frac{a_1' + \ldots + a'_n}{n}$$

$$\bar{\mu}_1' = \frac{\mu_1' + \ldots + \mu'_n}{n}$$

where the μ_i''s are the true values for the elements of the sample

For the sample random subsample of $n_1 < n$ elements, let

$$\bar{x}'' = \frac{x_1' + \ldots + x'_{n_1}}{n_1}$$

$$\bar{a}'' = \frac{a_1' + \ldots + a'_{n_1}}{n_1}$$

$$\bar{\mu}_1'' = \frac{\mu_1' + \ldots + \mu'_{n_1}}{n_1}$$

We suppose that values of $x_1', x_2', \ldots, x'_{n_1}$ are obtained from the respondents and that values of $\mu_1', \ldots, \mu'_{n_1}$ are obtained from records. (Sometimes, when more intensive reinterviewing is done, instead of the true values, $\mu_1', \ldots, \mu'_{n_1}$, values $y_1', \ldots, y'_{n_1}$ believed to have smaller biases than $x_1', \ldots, x'_{n_1}$ will be used. This type of analysis will be reported on later)

Let us define $\bar{z}'$ by the equation

$$\bar{z}' = \bar{x}' - (\bar{x}'' - \mu'')$$

Then $\bar{z}'$ is an unbiased estimate of $\bar{\mu}$, i.e.

$$E\,\bar{z}' = \bar{\mu}$$

Let $E\,x_i = a_i$; $\sigma^2_{x_i} = \sigma_i^2$; $\sigma_{x_i x_j} = \sigma_{ij}$

where

$$\sigma_{x_i x_j} = E\,(x_i - a_i)\,(x_j - a_j)\ .$$

Hence we assume that the response deviations of different respondents may be correlated.

Then

$$E\,(\bar{z}' - \bar{\mu})^2 = E\left[(\bar{x}' - \bar{a}') - (\bar{x}'' - \bar{a}'')\right]^2 + E\left[(\bar{a}' - \bar{a}'') - (\bar{\mu}'' - \bar{\mu})\right]^2$$

Also

$$E(\bar{x}' - \bar{a}')^2 = \frac{\sigma_w^2}{n}\{1 + (n-1)\,\rho\}$$

where

$$\sigma_w^2 = \frac{1}{N}\sum_{i=1}^{N} \sigma_i^2$$

and

$$\rho\sigma_w^2 = \frac{1}{N(N-1)} \sum_{i \neq j} \sigma_{ij}\ .$$

The response variance contribution is:

$$E\left[(\bar{x}' - \bar{a}') - (\bar{x}'' - \bar{a}'')\right]^2 = \frac{n-n_1}{nn_1}\,\sigma_w^2\,(1-\rho)$$

Also

$$E\left[(\bar{a}' - \bar{a}'') + (\bar{\mu}'' - \mu)\right]^2 = \frac{n-n_1}{nn_1}\,S^2_{a-\mu} + \sigma^2_{\bar{\mu}'}$$

where

$$S^2_{a-\mu} = \frac{1}{N-1}\sum_{i=1}^{N}\left[(a_i - \mu_i) - (\bar{a} - \bar{\mu})\right]^2\ .$$

Thus, finally,

$$\sigma^2_{\bar{z}'} = \sigma^2_{\bar{\mu}'} + \frac{n-n_1}{nn_1}\,\sigma_w^2\,(1-\rho) + \frac{n-n_1}{nn_1}\,S^2_{a-\mu}$$

where the first term, $\sigma^2_{\bar{\mu}'}$, is the term that would occur if there were no response error, the second term results from response variance, and the third term results from the sample response bias.

If no subsample is selected, then

$$E(\bar{x}' - \bar{\mu})^2 = \sigma^2_{\bar{a}'} + \frac{\sigma_w^2}{n}\left[1 + (n-1)\,\rho\right] + (\bar{a} - \bar{\mu})^2$$

where $\sigma^2_{\bar{a}'} = \frac{N-n}{Nn}\,S^2_a$ and $S^2_a = \frac{1}{N-1}\sum_{i=1}^{N}(a_i - \bar{a})^2$.

If, to simplify we put

$$\rho = 0$$

$$\sigma^2_{\bar{a}'} = \sigma^2_{\bar{\mu}'}$$

then

$$\Delta = E\,(\bar{x}' - \bar{\mu})^2 - E(\bar{z}' - \bar{\mu})^2 = (\bar{a} - \bar{\mu})^2 - \frac{n-n_1}{nn_1}\,S^2_{a-\mu} - \frac{\sigma_w^2}{n_1} \simeq (\bar{a} - \bar{\mu})^2\left(1 + \frac{n-n_1}{nn_1}\right) - \frac{n-n_1}{nn_1}\,T_{a-\mu} - \frac{\sigma_w^2}{n_1}$$

where

$$T_{a-\mu} = \frac{1}{N}\sum_{i=1}^{N}(a_i - \mu_i)^2\ .$$

If $\bar{a}-\bar{\mu}$ is small then Δ is small or negative, and clearly there is little or nothing to gain from the subsampling technique. Indeed, if $\bar{a}-\bar{\mu}$ were not frequently large relative to $\bar{\mu}$, there would be no point in discussing response bias.

When $\bar{a}-\bar{\mu}$ is large, then unless $S^2_{a-\mu}$ is large relative to $(\bar{a}-\bar{\mu})^2$, it should not require a very large value of n_1 to make $\frac{n-n_1}{nn_1} S^2_{a-\mu} < .1\ (\bar{a}-\bar{\mu})^2$; even a smaller multiplier than .1 should not be difficult to attain. Thus the costs of a subsample interview may be quite large compared to that of a first stage interview and still the double sampling technique would be sufficient.

Obviously, these results also indicate that the use of design methods such as stratification and regression estimates to reduce $S^2_{a-\mu}$ will greatly increase the efficiency of the double sampling approach to reduce response bias. These will be discussed in the larger paper but they demonstrate the importance of discussing explanatory variables - a subject to which we now turn.

8. Explanatory Variables. Response error will often vary with other variables. People with low incomes tend to overreport income and people with high incomes tend to underreport i.e., the response error in reporting income is a function of the "true" income. People who see a physician frequently and recently, tend to report their illnesses better than people with the same illnesses who have seen a physician less frequently and not recently, i.e. the response error in reporting illness is a function of how often and how recently a physician was seen for the illness.

The study of the relationships of response error to the characteristics of the respondent, the interviewer, the questionnaire and the environment in which the interviewing occurred provides information of importance in

a. Understanding the data that the survey provides
b. increasing the ability to design future surveys, and,
c. improving the estimates resulting from the current survey (See Section 7.)

We shall refer to the variables to which response error is related as explanatory variables.

Explanatory variables can be classified, as in the preceding paragraph according to whether they are characteristics of the respondent, the interviewer, the questionnaire or the environment. When records or other sources outside the respondent or interviewer are used, then the explanatory variables may also be classified according to whether they are available:

a. For each member of the population
b. For various strata of the population
c. For each member of the sample from the interview
d. For members of a subsample from sources outside the interview
e. For members of a separate sample from sources outside the interview, the members of the separate sample also being interviewed

When the explanatory variables are available for each member of the population, or for various strata of the population, then the usual conditions of stratification, or stratification after initial sample selection occur.

When the explanatory variables are available from the survey itself then the double sampling procedures with stratification and regression become available. However, in such cases the explanatory variables may themselves be subject to response error; sometimes the explanatory variables may not be available from the interview. For example, the number of visits to a physician stated in medical records may be quite different from the number stated by the respondent, and better related to the error of response. There will also be information only available from the physician or medical records, and not available from the respondent. Explanatory variables not available from the interviews may be more closely related to response error than those available from the interview, but sometimes their usefulness is limited because the cost of obtaining data on them is so great that they are obtained only for the subsample (See Section 7).

When a separate survey is made to obtain information to compute an estimate relatively free of response error the variance of the estimate will be larger than would the estimate based on a subsample of equal size from the interview sample, and if made at a different time or under different conditions the response error may be quite different.

Let us now consider the explanatory variable in connection with double sampling and let us consider only the use of the explanatory variable for stratification purposes, leaving regression models for the fuller study.

The population now is assumed to consist of elements (i,j), $i=1,2,\ldots, H$, $j=1,2,\ldots, N_i$ and the response random variable for element (i,j) is x_{ij}.

Let $E(x_{ij}\ i,j) = a_{ij}$, $\sigma^2_{x_{ij}\ i,j} = \sigma^2_{ij}$, and let the "true value" for (i,j) be μ_{ij}.

To estimate

$$\bar{\mu} = \frac{1}{H} \sum_{i=1}^{H} \bar{\mu}_i$$

where

$$\bar{\mu} = \frac{1}{N_i} \sum_{j=1}^{N} \bar{\mu}_{ij} \quad ,$$

we can use

$$\bar{x}' = \frac{1}{n} \sum_{i,j \in \mathcal{S}} x'_{ij}$$

where $\mathcal{S}$ is a simple random sample of n elements selected from the N ($N = \sum_{i=1}^{H} N_i$) elements of the population.

The sample, $\mathcal{S}$, is stratified and we denote the number of elements selected from the ith stratum by n_i, and assume $n_i > 0$, i=1,2,..., H.

From the n_i elements in $\mathcal{S}_i$, where $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_H$ is a partition of $\mathcal{S}$, we select m_i elements by simple random sampling and for these m_i elements we obtain the μ_{ij}.

Let

$$\bar{z}' = \sum_{i=1}^{H} \frac{n_i}{n} \left[\bar{x}'_i - (\bar{x}''_i - \bar{\mu}''_i) \right]$$

where $\bar{x}'_i$ is the mean of x for $\mathcal{S}_i$; $\bar{x}''_i$ is the mean of x for the subsample of $\mathcal{S}_i$ and $\bar{\mu}''_i$ is the mean of μ for the subsample of $\mathcal{S}_i$, i=1,2,...,H

Then $\bar{z}'$ is unbiased, and

$$\sigma^2_{\bar{z}} = \sigma^2_{E(\bar{z}'\ (n))} + E\, \sigma^2_{\bar{z}'|(n)}$$

where $(n) = (n_1, n_2, \ldots, n_H)$

It can be shown that

$$\sigma^2_{\bar{z}'} = \sigma^2_{\bar{\mu}'} + \sum_{i=1}^{H} E\left(\frac{n_i}{n}\right)^2 \left(\frac{n_i - m_i}{n_i m_i}\right)\left(\sigma^2_{w_i}(1-\rho_i) + S^2_{a_i-\mu_i}\right)$$

where

$$\bar{a}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} a_{ij}\ ,\quad \bar{\mu}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \mu_{ij}\ ,$$

$$\sigma^2_{\bar{\mu}'} = \frac{N-n}{N\,n} S^2_{\mu},\quad S^2_{\mu} = \frac{1}{N-1} \sum_{i=1}^{H} \sum_{j=1}^{N_i} (\mu_{ij} - \bar{\mu})^2$$

$$\sigma^2_{w_i} = \frac{1}{N_i} \sum_{j=1}^{N_i} \sigma^2_{ij}$$

$$\rho_i\, \sigma^2_{w_i} = \frac{1}{N_i(N_i-1)} \sum_{j \neq k} E\left[(x_{ij}-a_{ij})(x_{ik}-a_{ik}) \middle| i,j,k \right]$$

and

$$S^2_{a_i-\mu_i} = \frac{1}{N_i-1} \sum_{j=1}^{N_i} \left[(a_{ij}-\mu_{ij}) - (\bar{a}_i - \bar{\mu}_i) \right]^2$$

Then, if the m_i are fixed, it follows that

$$\sigma^2_{\bar{z}'} \sim \sigma^2_{\bar{\mu}'} + \sum_{i=1}^{H} \frac{N_i}{N} \left[\frac{N_i}{m_i N} - \frac{1}{n} \right] \left(\sigma^2_{w_i}(1-\rho_i) + S^2_{a_i-\mu_i} \right)$$

while if $m_i = p\, n_i$ then

$$\sigma^2_{\bar{z}'} = \sigma^2_{\bar{\mu}'} + \frac{1-p}{pn} \sum_{i=1}^{H} \frac{N_i}{N} \left[\sigma^2_{w_i}(1-\rho_i) + S^2_{a_i-\mu_i} \right]$$

If no subsample is selected, then as in Section 7,

$$E(\bar{x}'-\bar{\mu})^2 = \sigma^2_{\bar{a}'} + \frac{\sigma^2_w}{n} \left[1 + (n-1)\rho \right] + (\bar{a} - \bar{\mu})^2$$

Now the decision on whether to use the subsampling method depends largely on the relative sizes of $(\bar{a}-\bar{\mu})^2$ and

$$\frac{1-p}{pn} \sum_{i=1}^{H} \frac{N_i}{N} S^2_{a_i-\mu_i}$$

where we are omitting the terms in σ^2_w and $\sigma^2_{w_i}$

To measure the effectiveness of the explanatory variables in general, not only if the double sampling method is used, we suggest the ratio

$$\frac{\sum_{i=1}^{H} \frac{N_i}{N} (\bar{a}_i - \bar{\mu}_i)^2}{S^2_{a-\mu}}$$

i.e. the variance between strata divided by

the total variance of the individual bias, or the equivalent ratio

$$\frac{\sum_{i=1}^{H} \frac{N_i}{N} S^2_{a_i-\mu_i}}{S^2_{a-\mu}}$$

(We ignore the ratios $(N_i-1)\ N_i$.)

Similarly the effectiveness of explanatory variables in reducing bias when regression estimates are used should be measured by the ratio of the variance of $a-\mu$ about its regression line in terms of the explanatory variables.

Similar remarks hold for ratio estimates and for multiple explanatory variables using difference, regression or ratio estimates.

The use of double sampling with stratification will often lead to the desirability of using optimum strata. Considerable steps in that direction may be made from a general knowledge of the relation of the bias and explanatory variables, such as would result from earlier studies. The high cost of obtaining the μ_i will often justify multistage or sequential obtaining of the μ_i with improvements in the stratification at various stages of selection.

9. Response models. As has been mentioned earlier, the response to an interviewer's question is a choice made by the respondent. The mechanism governing that choice is relevant to the understanding of the data and the possibilities for reducing response bias.

We shall consider several cases.

Suppose that a question has q possible replies $y_1, y_2, \ldots, y_q$. Suppose that in the population there are N_i persons for whom the correct response is $y_i, i=1,2,\ldots,q$, $N_1+N_2+\ldots+N_q=N$.

Suppose that the response model has the following characteristics. Of the N_i persons for whom the correct response is y_i, M_i will give that response if selected, and the remaining N_i-M_i would choose one of all the possible responses with equal probability, i.e. with probability $1/q$.

Let u_i' be the proportion in the sample whose response is y_i . Then

$$u_i' = \frac{m_i + \frac{n-m}{q}}{n} = \frac{m_i}{n} + \frac{n-m}{nq}$$

where m_i is the number selected from the M_i whose correct response is y_i and give it, $i=1, 2,\ldots,q$ and $m=m_1+\ldots+m_q$. Now

$$* \qquad E\,u_i' = \frac{M_i}{N} + \frac{N-M}{Nq}$$

where $M = M_1+\ldots+M_q$ and hence u_i' is a biased estimate of $\frac{N_i}{N}$

$$E\,u_i' = \frac{N_i}{N} + \frac{\left(M_i - \frac{M}{q}\right) - \left(N_i - \frac{N}{q}\right)}{N}$$

If the same proportion of guessing takes place for each response then $M_i = t\,N_i$, say, and

$$E\,u_i' = \frac{N_i}{N} - \frac{(1-t)\left(N_i - \frac{N}{q}\right)}{N}$$

which shows that if $\frac{N_i}{N} > \frac{1}{q}$ then u'_i has a downward bias and that if $\frac{N_i}{N} < \frac{1}{q}$ then u_i' has an upward bias. From * it follows that unbiased estimates of $M_i,\ldots, M_q$ can be obtained if an unbiased estimate of the total proportion guessing, $1-\frac{M}{N}$, can be obtained in a supplementary interview or through other means. If the proportion, t, guessing is constant and an estimate of t can be obtained then, the ratio estimate

$$r_i' = \frac{1}{t}\left(u_i' - \frac{1}{q}\right) + \frac{1}{q}$$

will approximate $\frac{N_i}{N}$

To compute the variance of r_i' and to compute estimates of the mean

$$\sum_{i=1}^{q} \frac{N_i}{N}\, y_i$$

and the variance of the estimate under the restrictive assumption is not difficult but is left to later publication.

Estimates of N_i/N without the assumption that the proportions guessing are equal for all i, appear to require supplementary estimates of the M_i/N rather than only M/N.

Clearly many variations of this "guessing model" are possible. All are subsets of a more general model which can be formulated as follows:

The elements of a population may be classified into H classes. In the h^{th} of the H classes, there are M_h elements. Designate by $y_1, y_2, \ldots, y_q$ the possible responses to a question. Then there are H response matrices $P_1, P_2, \ldots, P_H$,

$$P_h = (p_{ijh}) \ , \ i,j=1,\ldots,q; \ h=1,\ldots,H .$$

If an element is in class h, and his "true response" is y_i, then p_{ijh} is the probability that he will actually respond y_j.

In the preceding case, H=2; the matrix P_1 is the identity matrix of q rows and q columns and the matrix P_2 is also a square matrix all of whose elements equal 1/q. M_1, the number of elements in the first class is M , the number of persons that do not guess, and, M_2 the number of elements in the second class is N-M the number who do guess.

Obviously, the general model need not include square matrices, or can handle them by putting certain of the p_{ijh} equal to 0. Furthermore, many variants on the models can be studied; for example, the main diagonal of P_2 might be 0 and the remaining elements all equal to 1/q-1.

A second class of models may be associated to the name "learning theory". The possibilities are many and only the simplest case can be discussed here because of the length of the paper.

Perhaps it will be useful to begin by considering the following case:

A person is asked to report his illnesses. It is well known that, in general, the greater the number of times he has seen a physician for an illness the higher the probability that he will report the illness.

If the person has seen the physician h times for an illness, denote by p_h the probability that he will report the illness, and hence $1-p_h$ is the probability that he will not report the illness.

If access to medical records is available then, by use of the double sampling procedure previously discussed, it is possible to obtain estimates $p''_1, p''_2, \ldots, p''_H$.

If x'_h is the number of cases of the illness reported by persons who made h visits in the larger sample and if on the basis of the subsample it is found that the proportion in the subsample, falsely reporting the condition is t''_h, then

$$x'' = \sum_{h=1}^{H} x'_h \cdot \frac{t''_h}{p''_h}$$

will approximate the "true number" having the illness. The mean square error of x" about the true value can be computed without great difficulty.

Let us suppose that the probability that the respondent who has made h visits to the physician became aware of the illness on visit g but not before, is

$$(1-p)^{g-1} \ p, \ g=1,2,\ldots$$

Then the probability that the person who makes h visits to the physician will know his condition is

$$* \quad p_h = 1 - (1-p)^h \ , \quad h=1,2,\ldots, H$$

where p is the probability that a visit results in the respondent learning of the condition.

To estimate p will require a smaller sample than to estimate $p_1, p_2, \ldots, p_H$. Furthermore as a means of guiding improvements in questionnaires and interviewing, the fact that a learning model may be approximately correct would be very useful.

To extend the model summarized by * to include the effects of time since the last visit to the physician would be important in discussing response errors in reporting illnesses; we are here more interested in indicating the use of learning and other behavioral science models as a means of understanding and reducing response error.

This paper, already too long, is part of a larger study on response error which is intended to deal with various aspects of response error measurement, understanding and reduction. The basic approach consists in formulating survey models that include response choice models in order to be able rationally to decide how to allocate the resources of a survey. In so doing the suggestion is made that even artificially simple models for the conclusions, decisions or actions to be taking on the basis of a survey may be adequate for guiding the design of the survey if used with judgment - even though these models are not actually used in the analysis of the survey when made.