Journal of Official Statistics
Vol. 5. No. 1, 1989, pp. 41–55

Early Survey Models and Their Use in Survey Quality Work

Gösta Forsman[1]

Abstract: There have been great advances in sampling models over the past 60 years. As these models have been developed, so has an awareness of the problem of nonsampling errors in surveys. Two lines have emerged in this work, namely (i) the development of theory and methods for handling specific sources of nonsampling errors, and (ii) the development of a comprehensive theory of an integrated treatment of survey errors. The latter line is characterized by the use of survey models. This paper deals with the early research on survey models up to the early 1970s, and looks at the application of these models in survey quality work.

Key words: Nonsampling errors; survey models; survey quality.

1. Introduction

At a statistical agency, survey quality work includes a variety of procedures such as evaluation studies, preventive control, and production control. One fundamental part of this work is the measurement of survey errors. Measurement studies provide information about quality that is useful for both the producer of the data and the user. The survey methodologist needs data on survey quality to improve methods and to allocate resources more effectively. The user of the statistics needs quality data to determine whether the survey estimates are reliable enough to meet his/her needs.

The early development of survey theory focused on the measurement and control of specific error sources. An important example is the very successful research on sampling errors. Since the 1930s there has been an increasing awareness of the problems of nonsampling er-

[1]Senior statistician, Statistical Research Unit, Statistics Sweden, S-115 81 Stockholm, Sweden. This paper is adapted from Forsman (1987).

Acknowledgements: I would like to thank the following persons who kindly supplied information. C. Andersson, B. A. Bailar, S. Berg, P. P. Biemer, T. Dalenius, J. C. Deville, L. Fabbris, I. Fellegi, G. B. Gray, M. Hansen, W. Keller, L. Kish, G. Koch, J. Koop, J. T. Lessler, I. Lyberg, L. Lyberg, R. Platek, J. N. K. Rao, M. Ribe, H. Strecker, P. V. Sukhatme, G. Theodore, A. Sunter, V. Verma, and S. S. Zarkovich. The editor and two anonymous referees provided many helpful suggestions. I am, however, responsible for any defects that might be found in the paper.

rors. India and the United States led the early development in this field. Forerunners were the Indian Statistical Institute, led by Mahalanobis, the Indian Council of Agricultural Research, where P. V. Sukhatme worked, and the U. S. Bureau of the Census, where Hansen, Hurwitz, Tepping, Madow, and others were pioneers. Contributions from the United Kingdom were also important. At the Rothamstead Experimental Station, Fisher, Yates, Cochran, and others did research on statistical experiments in the 1920s and 1930s. This work had a strong influence on the development of survey theory.

Studies of specific error sources have continued to be an important part of survey quality work. In the 1940s a parallel development emerged that aimed at an integrated control of all sources of errors and thus of the total error. In this research, what is called mixed error models were developed; later, the term survey models has been widely used.

We can distinguish three fields of application for survey models:

1. As already indicated, a survey model allows an integrated treatment of various error sources. Thus using the model, the total error can be estimated. Note that the total error given the model is the error resulting from the error sources that the model takes into account. The "real" total error of a survey estimate may be affected also by other sources of error.

2. Survey models can be used to estimate the relative impact of different error sources on the total error. For recurrent surveys, this allows a reallocation (if necessary) of resources to effectively control the error sources.

3. Survey models might also be applied to a specific source of error to study the magnitude of its components. For example, if applied to the response error, we can estimate the total response error and its components, the response bias and the response variance. This also can lead to an improved allocation of the resources among survey operations.

We will review the early development of survey models up to the early 1970s and discuss their use in the subsequent work on survey quality. It is mainly these models and modified versions of them that have been used in survey practice so far. The presentation is restricted to models that include estimation procedures for the error components. The theoretical development is reviewed in Section 2. Section 3 contains examples from survey practice and in Section 4 we discuss the application, or lack of application of survey models.

2. Survey Models

2.1. Models for variable measurement errors

The work to develop survey models (led by Indian and American statisticians) concentrated on sampling variance and measurement variability. Two important sources of measurement variability were identified early:

a) the error that depends on the tendency of the interviewers (or enumerators or observers, depending on the data collection mode) to affect the respondent's answers, and

b) the error that emerges from the fact that the answers to a question can be different if the same respondent is asked the same question on different occasions.

There has not yet emerged a common nomenclature for these sources of error. In this section, I refer to them as interviewer error and respondent error, respectively. It should be noted that the respondent error includes some of the effect of the interviewer error if different interviewers ask the same questions on the two occasions.

In the United States, Rice (1929) showed that the interviewers' own attitudes affected the respondents and could lead to a response error. There was an urgent need to measure this type of error at the U.S. Census Bureau. The data collection in the Decennial Censuses of Population and Housing was conducted by thousands of temporarily employed interviewers whose skills could vary considerably. This type of error was also well-known in India, e.g., in the crop surveys where the observers might classify the same field very differently.

The modeling of the interviewer error was done somewhat differently in different agencies. Central to the U.S. Census Bureau's model was that each interviewer generated clusters of responses. Then, by allocating a random subsample to each interviewer, the error could be measured by a cluster sampling (variance) formula. The situation here differs, however, from that in cluster sampling. In cluster sampling, the correlation seen in the data reflects the correlation that exists in the population. In the U.S. Census Bureau's model, on the other hand, the correlation is a result of the observation and data collection process. This error component was called the correlated response variance. In India and in some other agencies in the United States, this error was regarded as a bias due to the interviewer (or the observer). The interviewers were considered a simple random sample from a population of interviewers. The variance among the biases associated with these interviewers in the population was often called the interviewer variance. This variance component could then be estimated from the sample of interviewers. The interviewer variance is often regarded as identical to the correlated component of the response variance according to the Census Bureau model. This is only approximately true, however, since, theoretically, the correlated component of the response variance can take on negative values.

The modeling of the respondent error was also done differently at different agencies. At the U.S. Census Bureau the random nature of this error was discussed early (see Palmer (1943), and Deming (1944)). In the model developed at the Census Bureau it was assumed that an answer to an interview question is generated by a random process. As a consequence – even a response from a given respondent to a given interviewer has a probability distribution. A similar situation is assumed in Sukhatme (1954) and Sukhatme and Seth (1952) and probably also in Mahalanobis (1946), although Mahalanobis does not explicitly describe a survey model as we have defined it here. Another way of modeling the respondent error is to assume that only one answer is possible for each respondent-interviewer-question combination. A given respondent could, however, provide different answers to the same question to different interviewers. The stochastic element in a survey model with this assumption is entirely due to the sampling processes and the allocation of respondents to interviewers. Both interviewers and respondents are usually regarded as sampled from large populations. One can interpret the survey model described by Stock and Hochstim (1951) as based on this deterministic approach. The same goes for the later model by Murthy (1967) and the conceptual discussion in Zarkovich (1966).

The models for the total survey error, which considered the sampling error and the two variable measurement error types described above, were usually formulated according to two basic ideas. The Census Bureau used a mean square error decomposition approach founded in sampling theory, while other agencies used a linear model approach founded in the analysis of variance (ANOVA) technique.

2.1.1. The mean square error decomposition approach

The Census Bureau model assumes a set of general conditions under which the survey is conducted. The survey is regarded as one trial from among a large set of conceived repetitions of the survey under the same general conditions. This means that a measurement derived from the survey has a well-defined, but unknown, probability distribution. The model postulates the existence of a true value, x, for each sampling unit. We denote the measurement for the ith element at the tth trial by y_{it}. Now, the conditional expected value of y_{it} over all possible samples that include the ith element and all possible trials that have resulted in such a sample, is

$$E(y_{it}|i) = Y_i. \tag{2.1}$$

The difference between the observation on the ith unit in a particular survey and the conditional expected value of that unit is

$$d_{it} = y_{it} - Y_i.$$

d_{it} is called the response deviation.

Assume now, that in a specific trial, t, the population mean, $\bar{X}$, is to be estimated by $\bar{y}_t$, the sample mean from a simple random sample of n units. Then the total error $\bar{y}_t - \bar{X}$ is measured by MSE($\bar{y}_t$), which can be decomposed as:

$$\text{MSE}(\bar{y}_t) = \sigma_S^2/n + \sigma_R^2[1 + (n-1)\varrho]/n + 2(n-1)\sigma_{RS}/n + B^2. \tag{2.2}$$

In (2.2), the first term is the sampling variance of $\bar{y}_t$, defined as the variance among the Y_i-values in the population, divided by n. The second term is the response variance, defined as the variance of $\bar{d}_t$, the average of the response deviations for the sample. This term can be further decomposed into the simple response variance, σ_R^2/n, (which is the error component corresponding to the respondent error) and the correlated response variance, $\varrho(n-1)\ \sigma_R^2/n$. Here, ϱ is the intraclass correlation coefficient among the response deviations for a trial (survey), defined as

$$\varrho = E(d_{it}, d_{i't})/\sigma_R^2, \quad i \neq i'.$$

It is important to recall that the sampling variance measures variations caused by the sampling process, while the response variance measures variations assumed to characterize the measurement operation. The third term in (2.2) is the covariance of the response and sampling deviations, which is normally regarded as very small – it is zero for a complete census. The fourth term, finally, is the squared bias.

An important feature of the model is its broad applicability. It may be applied to any sequence of survey operations, i.e., either the full sequence or a subset of operations (for instance, interviewing and coding). Applied to the full sequence, the response variance reflects contributions from all operations such as interviewing, coding, editing, and so forth. Applied to coding alone, the response variance reflects only coding and the response variance becomes a coding variance. Analogously to (2.2), coding gives a contribution to the MSE of the form

$$\sigma_C^2[1 + (n-1)\varrho_C]/n + B_C^2. \tag{2.3}$$

For surveys with interviewers, the correlated response variance may be especially large. It is then important to note that this component does not decrease when the number of sampled units within an interviewer's assignment increases. Hence a relatively low value of ϱ can have a considerable effect on the total response variance and also on the total MSE. This is readily seen if the correlated response variance is assumed to depend entirely on the interviewers (i.e., on the "interviewer error" described in Section 2.1). The response variance is then given by

$$\sigma_R^2[1 + \varrho(m-1)]/n, \tag{2.4}$$

where m is the (average) number of respondents assigned to an interviewer.

The Census Bureau survey model was first presented in Hansen, Hurwitz, Marks, and Mauldin (1951). In this paper, Hansen et al.

assumed that the correlated response variance depends entirely on the interviewers. They showed that the correlated component of the response variance could be estimated by means of interpenetrating subsamples. They also showed that the common textbook estimator of the sampling variance of $\bar{y}_t$ actually estimated the sum of the sampling and simple response variances. During the 1950s, the model was further elaborated on and eventually presented in the two widely recognized papers by Hansen, Hurwitz, and Bershad (1961) and Hansen, Hurwitz, and Pritzker (1964). In these papers, the correlated component of the response variance was defined as dependent not only on the interviewers, but on all field personnel. The correlation between answers to different interviewers was permitted to be nonzero, reflecting the possible correlations arising from supervisors, coders, editors, keyers, etc. The above assumptions apply to (2.2).

In Hansen et al. (1964), an estimator of the simple response variance for 0,1-variables was presented. This estimator was derived under the assumption that independent repeated measurements of the sampled units were conducted. In this case the original survey and its replication (the reinterview) were assumed to be two independent randomly selected trials. Now the gross difference rate,

$$g = \sum_{i=1}^{n} (y_{i1} - y_{i2})^2/n, \tag{2.5}$$

divided by two, can be shown to be an unbiased estimate of σ_R^2/n. Hansen et al. (1964) defined an index of inconsistency as the ratio of the simple response variance to the total variance of individual responses, σ_y^2, that is

$$I = \sigma_R^2/\sigma_y^2. \tag{2.6}$$

For a Bernoulli random variable with parameter P, the total variance σ_y^2 is $P(1-P)$. An estimator of the numerator is then $g/2$. The denominator may be estimated by $\bar{y}_t(1-\bar{y}_t)$, $t = 1, 2$. Here $\bar{y}_t$ is the proportion of sample units that belongs to the category of study in trial t data from either trial 1, or trial 2, or from both trials may be used). Obviously, the index takes values between 0 and 1. Low values of the index indicate that the measurement process is under control.

The two procedures for estimating error components mentioned above, interpenetration and repeated measurements, are the basic methods available in the estimation process. Fellegi (1964) demonstrated how the two procedures could be combined. In his notation, the assignment of the jth interviewer is $S_{j(1)}$, $S_{j(2)}$, $j = 1, \ldots, k$, where $S_{j(1)}$ and $S_{j(2)}$ are randomly allocated assignments for the jth interviewer in the original and reinterview surveys, respectively. $S_{j(1)}$ and $S_{j(2)}$ are not the same for a given interviewer. The model is similar to the Hansen, Hurwitz, and Pritzker (1964) model but differs in that the conditional expected values of a measured value y_{ijt} given a respondent, i, and an interviewer, j, over the trials need not be the same for the original survey and the reinterview. Fellegi's data collection design accomodates the definition of several types of correlation among the response deviations.

Bailar and Dalenius (1969) demonstrate the potential usefulness of the procedures interpenetration and repeated measurements. The procedures are reviewed (also in combination) in several basic study schemes aiming at estimating different variance components in the Census Bureau model. The study schemes are classified according to repetition and interpenetration in two dimensions, called the sample dimension and the trial dimension. Repetition in both dimensions is characterized by the use of the same sample and the same field personnel (e.g., interviewers and coders) in a replicated study, repetition in the sample dimension combined with interpenetration in the trial dimension implies the use of the same sample and different field workers in a replicated study, etc. Some of the study schemes are even more sophisticated than Fellegi's, but these schemes are also more difficult to implement.

Sometimes, when interpenetration or repetition cannot be applied in a survey, data from other surveys are used instead. For example, data from a match between a labor force survey and a census with labor force items may sometimes be analyzed as if they were reinterview data. Moreover, data from two independently conducted surveys with similar questions and data collection procedures on the same population may be treated as if interpenetration had in fact been applied (see Tepping and Boland (1972)).

Repeated measurements can also be used to estimate the bias component. The measurements must, however, be carried out with a preferred procedure that can be assumed to provide data close to the true values.

Murthy (1967) and Des Raj (1968) present variance decomposition models. They both arrive at expressions of the variance of the sample mean that are similar to the Census Bureau model decomposition, although their components have different definitions. Murthy conceives of the survey as having two steps of randomization:

i. a sample of population elements, s, and

ii. a sample of survey personnel, r.

He defines y_{ij} as the value obtained by the jth interviewer for the ith element. Since Murthy assumes the deterministic response model described in 2.1, this value is not a random variable. He gives the following expression for the variance of the sample mean, $\bar{y}$:

$$V(\bar{y}) = \sigma_s^2/n + \sigma_d^2[1 + (m-1)\,\varrho]/n,$$

where the terms are called sampling variance, simple or uncorrelated response variance, and correlated response variance, respectively. The names are the same as the names of the components of the Census Bureau model, but they are not the same components; the response deviations are defined differently.

Contrary to Murthy, Des Raj defines the observed value y_{ij} as a random variable. In his design the interviewers are allocated randomly to primary sampling units which have been selected with probabilities proportional to size. Thus, Des Raj is the first to present a survey model based on a PPS sampling design.

2.1.2. The linear model approach

Linear survey models may also be constructed in many different ways. The models emerged from the analysis of variance theory. In the late 1930s and early 1940s, the Indian Statistical Institute was the first to use ANOVA-type models in survey practice. Under the leadership of P.C. Mahalanobis sampling designs for crop surveys with embedded experiments based on interpenetrating subsamples were developed. The purpose of these experiments was to control the individual investigator bias that had been encountered in surveys where different investigators had unknowingly been allotted the same fields.

One of the first examples in the literature of measuring the overall error by means of linear survey models was provided by Stock and Hochstim (1951). The authors seem to assume the deterministic response model as mentioned above, i.e., only one answer is possible for each respondent-interviewer-question combination. In this application, the deterministic model presupposes a population of N respondents and a population of K interviewers. Then, assuming that each interviewer, j, interviews each respondent, i, in the population, the data generating process may be modeled in the following way:

$$y_{ij} = \bar{y}_{..} + I_j + e_{ij},\ i=1, .., N,\ j=1, .., K,$$

where

$\bar{y}_{..}$ is the mean of all $N \times K$ observations,

I_j is the individual interviewer bias for interviewer j, i.e., the difference between $\bar{y}_{..}$ and the mean of the N observations, $\bar{y}._j$, of interviewer j, and,

e_{ij} is the deviation between the observed value and $\bar{y}.. + I_j$ when interviewer j interviews respondent i.

It is assumed that no correlation is present between I and e. The survey design is such that an interpenetrated subsample, drawn by simple random sampling, is randomly allotted to each of k interviewers. The interviewers are sampled by simple random sampling from a population of K interviewers.

Now, if assuming that the sampling fractions are small, the variance of the sample mean, $\bar{y}$, is approximately:

$$V(\bar{y}) = \sigma_I^2/k + \sigma_e^2/n,$$

where n is the total sample size.

σ_I^2 is the variance among the I_js, defined by $\Sigma_j^K (\bar{y}._j - \bar{y}..)^2/(K\text{–}1)$. This is the interviewer variance mentioned in Section 2.1, and it is present in most ANOVA-type survey models (although it may be slightly differently defined). σ_e^2 is the variance between respondents within interviewers, averaged over all interviewers. σ_e^2/n is the sampling variance and would be the total variance under the model if there were no interviewer effects. Note that the model does not take into account the existence of true values.

Sukhatme (1954) presents a linear survey model different from the Stock and Hochstim model in that (i) the existence of true values is assumed, and (ii) one of the error components (and thus y_{ij}) is regarded a random variable. Like Stock and Hochstim, Sukhatme assumes finite populations of N respondents and K interviewers. He lets

$$y_{ij} = x_i + \alpha_j + \varepsilon_{ij},$$

where x_i is the true value. α_j is defined as "the bias of the jth enumerator in repeated observations on all units." α_j is very close – if not identical – to I_j in the Stock and Hochstim model. ε_{ij} is the random deviation of the reported value from $x_i + \alpha_j$. It is assumed that the ε_{ij}s are independently distributed with mean 0 and variance σ_ε^2 for all $i = 1, \ldots, N$ and $j = 1, \ldots, K$. Assuming the same sampling and measurement design as in the above description of the Stock and Hochstim model, Sukhatme derives the following expression for $V(\bar{y})$:

$$V(\bar{y}) = \sigma_x^2 (1/n - 1/N) + \sigma_\alpha^2 (1/k - 1/K) + \sigma_\varepsilon^2/n,$$

or, if N and K are large:

$$V(\bar{y}) \approx (\sigma_x^2 + \sigma_\varepsilon^2)/n + \sigma_\alpha^2/k$$

σ_x^2 is the variance among the true values in the population. It cannot be estimated separately from σ_ε^2 with Sukhatme's design. Sukhatme shows that the mean square between observations within enumerators is an estimator of the sum of σ_x^2 and σ_ε^2. This result is similar to the finding that the common sampling variance estimator estimates the sum of the sampling and simple response variances in the Census Bureau model. However, the components are differently defined in the two models. σ_α^2 is, analogously to σ_I^2 above, defined as $\Sigma_j^K (\alpha_j - \bar{\alpha})^2/(K-1)$, where $\bar{\alpha} = \Sigma_j^K \alpha_j/K$.

Sukhatme also derives the correlation, ϱ', between responses obtained by the same interviewer. If the finite population correction is small, the variance of a single observation is approximately $\sigma_y^2 = \sigma_x^2 + \sigma_\alpha^2 + \sigma_\varepsilon^2$. An approximate expression for ϱ' is

$$\varrho' = \sigma_\alpha^2/\sigma_y^2 = \sigma_\alpha^2/[\sigma_x^2 + \sigma_\alpha^2 + \sigma_\varepsilon^2],$$

which gives $V(y) = \sigma_y^2\,[1 + \varrho'((n/k) - 1)]/n$.

A similar model is presented by Kish (1962), who, however, pools the x and ε components. The reason for this is that (as is mentioned above) the variance components σ_x^2 and σ_ε^2 cannot be estimated separately with Kish's (or Sukhatme's) design since it does not include repeated measurements. Kish's model can be written

$$y_{ij} = y_{ij}' + \alpha_j,$$

where $y_{ij}' = x_i + \varepsilon_{ij}$. The definitions of the components seem to be the same as in the Sukhatme model. The intraclass correlation, ϱ^*, between responses obtained by the same interviewer is defined as:

$$\varrho^* = s_a^2/ (s_a^2 + s_b^2), \tag{2.7}$$

where s_a^2 and s_b^2 are the sample estimates of the variance between interviewers and within interviewers, respectively.

The Sukhatme (1954) model is a simplified version of the model presented in Sukhatme and Seth (1952). The latter model allows several observations on each unit and an interaction between the interviewer and the respondent.

The above linear models are based on the assumption that the expected values of the measurements do not depend on the sample. This assumption is usually not made in the variance decomposition models. E.g., the Census Bureau model takes the covariance between of the response and sampling deviations into account, see the third term in (2.2).

2.1.3. Other models for measurement variability

There are other approaches to developing survey models for surveys intended to estimate proportions in the presence of classification errors. In these models, the key concept is what is called misclassification probability. Assume, for example, that the true value, x_i, for unit i is 1 if i belongs to some category, say C, and 0 otherwise. Errors that give rise to the misclassification of a unit are considered. The survey procedure is assumed to generate a stochastic variable y_i, such that

$$\varphi = \Pr\ (y_i = 1 \mid x_i = 0), \text{ and}$$

$$\theta = \Pr\ (y_i = 0 \mid x_i = 1),\ i = 1, \ldots, N.$$

φ and θ are the misclassification probabilitites.

Casady (1966) derived two such survey models for the analysis of reinterview data in the Health Interview Survey conducted by the U.S. National Center for Health Statistics. Casady defined the "within element response variability" and an index of inconsistency and presented estimators of these parameters using the two models. The models differed in that the misclassification probabilities in the first model were assumed to be constant over different trials while the misclassification probabilities were permitted to vary between trials in the second model.

Swensson (1969) showed that the first Casady model could be regarded as a special case of the Census Bureau model (only the sampling variance and the simple response variance are regarded – the correlated response variance cannot be studied under the given definition of misclassification probabilities).

The misclassification approach was discussed in Cochran (1968), who defined the misclassification probabilities as dependent on the unit. Bailar and Biemer (1984) showed that the misclassification probabilities can be formulated as dependent on both the unit and the operator (i.e., an interviewer, coder, supervisor, etc.). This allows a correlated measurement error component to be estimated using the misclassification probability approach.

2.2. *Survey models for systematic errors*

Systematic errors are normally studied by comparing the survey data to preferred data. Higher quality data are usually obtained from reinterviews or record checks. The method allows an estimation of the bias term as it appears in, e.g., the Census Bureau model (see formula 2.2). This term has been extensively studied within the U.S. census evaluation programs.

In addition to the Census Bureau work on survey models, bias models were developed for specific survey situations. Kish and Lansing

(1954) developed a model for the case where not only the observed values but also preferred values, obtained from a preferred data collection procedure, are available. These preferred values were, however, not regarded as good as the true values. This model was to estimate the error in a study of the market value of houses, a study that was part of the 1950 Survey of Consumer Finances in the United States.

The well-known randomized response model presented by Warner (1965) can be regarded as a survey model since it takes into account different error sources. In Warner's model, response errors with known probabilities are intentionally introduced to eliminate nonresponse and erroneous answers to sensitive questions. This technique makes it possible to construct unbiased maximum likelihood estimators of population means and totals.

The connection between survey models and randomized response models is even more evident in the paper by Abul-Ela, Greenberg, and Horwitz (1967), who extended the Warner model to a trichotomous randomized response model. Contrary to the Warner model, the respondent is assumed to tell the truth with a probability that is allowed to be less than 1 in the Abul-Ela et al. model.

3. Use of the Early Survey Models

Several aspects of survey quality work have been affected by the use of early survey models. These models have been used in regular surveys, in evaluation studies, and in development work. In addition, the early survey models provided a conceptual framework that has completely permeated survey practice. The simple and correlated components of the response variance, the interviewer variance, and the index of inconsistency have become well-known and useful concepts, often theoretically discussed in technical reports even in cases where the models have not been explicitly applied.

In this section we will give concrete examples of work on data quality, guided and inspired by the early models presented above. The examples are confined to the models described in Sections 2.1.1–2.1.2, which are the most frequently used. It should be emphasized that the list is by no means a comprehensive review of the quality work guided and inspired by these models.

For each of the two main approaches for formulating survey models, the MSE decomposition approach and the linear model approach, we saw that different survey models may be developed. However, the two approaches can lead to very similar models which can be used for decomposing the total error into similar components, the same estimators of these components are used and, consequently, also the same data collection designs. The choice between different approaches is then probably more dependent on the way the statistician is used to structuring statistical problems than on other considerations. There is a difference in the use of the approaches. In the quality work performed by governmental agencies, the mean square error decomposition approach is by far the most common. Probably, the large-scale work with these models at the U.S. Bureau of the Census and Statistics Canada has influenced this decision. The linear models are more frequently used in research work on interviewer effects conducted in survey agencies outside the national bureaus.

3.1. *Models based on mean square error decomposition*

Not very surprisingly, the most extensive quality work based on survey models has been conducted at the U.S. Bureau of the Census. The bulk of the work has been done within two of the Bureau's major projects: the Decennial Census of Population and Housing and the Current Population Survey.

A continuing program of research, evaluation, and experimental studies has been conducted as a part of the censuses and during the intercensal periods. The results of the 1950 census experiments led to important changes in procedures adopted for the 1960 census. In one of these experiments, a set of interviewer-assignment areas was designated. In these areas, the interviewers' assignments were randomly allocated according to the design postulated in the Hansen et al. (1951) model. This experiment dealt with the variance between and within interviewers. The intraclass correlation of response errors within interviewers was also estimated. In U.S. Bureau of the Census (1985), these intraclass correlations are reported for items such as race, age, educational attainment, income, etc. Among the items having the largest ϱ's were the not-reported-categories indicating the influence of enumerators on item nonresponse rates.

For items that are typically difficult to measure (i.e., occupation, education, and income) the correlation was often around .03 (see also Hansen and Tepping, 1969, p. 11). This seems small, but when the average size of an interviewer's assignment is about 700, the factor $[1 + \varrho\ (m-1)]$ in (2.4) becomes larger than 20, leading to a substantial contribution to the total variance even for a moderate σ_R^2. These and similar findings showed that the variability in the complete census results was as large as if only a 25% sample had been taken (in the absence of interviewer effects). This was true even for areas with populations smaller than 5000 people. These findings along with studies of the bias and experimental studies of self-enumeration, etc., led to the following procedural changes for the most difficult items to measure in the 1960 census:

i. The data collection was based on a 25% sample.

ii. A self-enumeration procedure was introduced for this sample.

The interviewers were, however, still engaged in the data collection for the 1960 census. Interviewers delivered the questionnaires to the households and completed them for those households that did not mail in a completed form or whose questionnaires were inconsistent. This led to an interviewer influence on the variance in the 1960 census too. It was much smaller than in the 1950 census, but, however, still important for a number of items.

In the 1970 and 1980 censuses, changes were made in the census-taking procedures in that the questionnaires were delivered by mail to most of the population (95% in the 1980 census). The enumerators still had an important role in the follow-up procedures, and enumerator variance studies were also made in the evaluation programs of these censuses.

Within all content evaluation programs of the censuses from 1950 to 1980, large-scale reenumeration studies were conducted to obtain estimates of response variance and bias. The reenumerations were conducted as reinterviews or as a record match to the Current Population Survey.

The Census Bureau model was also applied to the coding process, as described above in formula (2.3). Jabine and Tepping (1973) presented estimates on the simple and correlated coding variance components (presented as relvariances) for 1960 census data. These were related to sampling and total response relvariances, as well as to response and coder bias (the latter was based on 1970 census data).

In the Current Population Survey (CPS) a continuing reinterview program has been conducted since the early 1950s. These studies are primarily designed to control the field procedures, rather than measuring the simple response variance according to the Census Bureau model. Nevertheless, reinterview data are continually used to derive the index of inconsistency for various items. According to U.S. Bureau of the Census (1978), this measure has an important role in CPS quality work: "The index is used primarily to monitor the measurement procedures over time. Substantial changes in the indexes that persist for several months result in review of field procedures to determine and remedy the cause."

Experiments aiming at measuring the correlated components of the response variance are not conducted in the CPS. Tepping and Boland (1972) report, however, from a study where data from the Monthly Labor Survey, carried out during six months in 1966 concurrently with the CPS, provided estimates independent of the regular CPS estimates for several items. The two estimates could then be used for estimating the correlated response variance component. In this paper, Tepping and Boland present estimates of the ratio of the correlated response variance to the sum of the sampling variance and the simple response variance, i.e., in terms of Section 2.1:

$$(m-1)\ \varrho\ \sigma_R^2/(\sigma_R^2+\sigma_S^2).$$

The estimated ratios range between 0.5 and 1.0.

In Canada, the Fellegi model was applied in an experimental pilot study preceeding the 1961 Canadian Census of Population. The results were similar to those found in the U.S. census in that the correlated response variance, derived as the mean of the correlated response variances in the two surveys, was "several times as large as the simple response variance for all except the basic population counts, such as the number of males, sons, married persons, persons of certain age, etc." (Fellegi (1964). Fellegi concluded that, for most characteristics, "considerable gains in the total response variance may be made by reducing the size of the enumerators' assignments". Fellegi argued that the Canadian Census should use a self-enumeration procedure. To determine if such a procedure would increase the simple response variance, he compared the index of inconsistency for a self-enumeration survey with the index of inconsistency for an interview survey. He used items from the 1960 U.S. census (self-enumeration) that corresponded to items in his pilot study for the 1961 Canadian census (interviews). Fellegi found that the values of the simple response variance were rather similar despite the different procedures. As a result of these findings, the 1971 Population Census of Canada was substantially modified. Self-enumeration was introduced along with a sample based collection of most census questions. Later, Krotki and Hill (1978) compared the Fellegi estimates of the correlated response variance with the corresponding estimates from the 1971 and the 1976 Canadian censuses. They found that for almost all characteristics examined, the magnitude of the estimates were considerably reduced.

In Spain, an evaluation program of the General Population Survey (which includes, e.g., labor force items) has been conducted since the early 1970s. The program is based on 3 000 reinterviews each quarter. The purposes of the program are to control the work of the interviewers and to evaluate the general quality of the results. According to Sanchez-Crespo (1973, 1981), the quality evaluation is based on the U.S. Census Bureau model. In the 1981 paper, estimates of the total response variance, the simple response variance, and the correlated response variance are presented for the variable "unemployed" (the study design used for estimating the correlated component is, however, not described). The correlated component was found to give the largest contribution to the total response variance.

In Belgium, a variance decomposition model for surveys with reenumerations, developed by Strecker and Wiegert, was applied in the 1979 Census of Agriculture. The application was limited to one variable, viz., the number of pigs. A study was conducted which provided both replicated data on which estimates of the simple response variance were based, and preferred data, on which bias estimates were based. The impact of the simple response variance component was considerable, as the following example (from Strecker, Wiegert, and Kafka (1984)) shows. The mean square error was defined as the sum of the simple response variance, the sampling variance, and the squared bias. The relative MSE for the estimate of the mean number of pigs per holding was estimated to 4.61 %. If the relative MSE for this variable had been defined as the sum of the sampling variance and the squared bias only, it would have been 1.92 %. Thus the simple response variance more than doubled the relative MSE.

At Statistics Sweden, the quinquennial Censuses of Population have been evaluated during the last decades. An evaluation based on a survey model is conducted only for labor force items, though. This is based on two sets of data. One data set is created by a match between census labor force items and Labor Force Survey (LFS) data collected during the same time as the census is taken. The LFS data set is then regarded as an independent replication of the census data. The other data set is created by a reconciliation of the census-LFS match and is regarded as preferred data.

Lyberg (1986) reports estimated error components for the items "hours/week at work" and "outside the labor force" for the 1980 census data. In general, the simple response variance is small compared to the squared bias. This fact together with the assumption that the correlated response variance is small (because the census data are collected by mail) has led to the conclusion that bias is the major problem in the Swedish population census. However, neither the impact of the editing personnel on the estimates nor the impact of other items except those mentioned above have been studied.

3.2. *Linear survey models*

In the above mentioned papers by Stock and Hochstim (1951), Sukhatme and Seth (1952), and Kish (1962), examples of studies of interviewer effects are described. In recent years Kish's simple model has been frequently used. The parameter of study in these applications is the intra-class (or intra-interviewer) correlation, ϱ^*, defined by (2.7). The Kish approach is relatively undemanding in terms of experimental design and permits comparisons between studies involving different numbers of interviewers and respondents. We shall review two examples of such studies.

Collins (1980) reports three experiments on interviewer variability conducted by the Social and Community Planning Research (SCPR) in the United Kingdom. The experiments took place in Southampton, North Yorkshire, and Milton Keynes. The questionnaires dealt with the problems faced by the disabled, environmental preferences, and different aspects on living and working, respectively. The estimated interviewer effects (ϱ^*) were generally larger in the Southampton study than in the two other studies. One possible explanation for this is the topic of study. Some categories of question are more prone to interviewer variability than others. Examples of questions prone to interviewer variability would be questions which the interviewer is reluctant to ask and the answer is often imputed from responses given elsewhere in the interview. Such reluctance can be common in a study of disability and its consequences. The results from the North Yorkshire and the Milton Keynes experiments were remarkably similar, despite the fact that the former dealt mainly with attitudinal items and the latter mainly with factual items. This confirmed results from comparisons reported by Kish (1962), who could not find any systematic differences in ϱ^*-values between attitudinal and factual items.

At the Survey Research Center at the Institute for Social Research (ISR), University of Michigan, U.S.A., the Kish model has been frequently applied in measuring interviewer effects, first by Kish in studies of factory workers' job attitudes. In recent years, Groves and others have applied the Kish model in various telephone surveys. Groves and Magilavy (1986) reviewed nine ISR telephone surveys and the estimates of ϱ^* for 297 survey items. Other interview surveys were also reviewed in which similar models for interviewer effects were applied. The average values of ϱ^* were in eight of the nine ISR surveys under .01, but varied considerably between different statistics. The lowest average of ϱ^*, .0018, was found in the survey with the largest interviewer workload, which, together with other observations, led Groves and Magilavy to the interesting conclusion that the (Kish) survey model underlying ϱ^* might be further developed to reflect larger interviewer variability in the initial cases completed by the interviewers.

Like Collins and Kish in their studies, Groves and Magilavy did not find any evidence that factual items as a class are subject to different interviewer effects than are attitudinal questions. Groves and Magilavy also discuss two issues concerning interviewer effects which have been largely overlooked in the literature, namely, the stability of the estimates of ϱ^*, and the causes of interviewer effects.

4. Discussion

In this paper we have reviewed several important applications of early survey models. Some of them, like the U.S. Census Bureau model and the Kish model, have shown a broad applicability. They have been used not only within the agency for which they primarily were designed, but also in other contexts with quite different survey environments. Despite this, there is reason to ask why survey models have not been more extensively applied in survey quality work. After all, outside the United States, Canada, and perhaps some other countries, applications of survey models are rare. Applications do appear in certain experiments and surveys, but often as a result of a single researcher's interest in the field. These scattered applications often concern a small number of variables only. There are different reasons for this state of affairs.

i. The models do not cover all possible error sources. The survey models we reviewed in Section 2 mainly concern content errors and sampling errors. They do not account for, e.g., frame errors, coverage errors, and nonresponse errors.

ii. The models are based on assumptions that are seldom met in survey situations. For instance, when estimating the simple response variance, a common assumption is that reinterview responses are independent of the original answers and have the same distribution. Bailar and Dalenius (1969) showed that the simple response variance component could be estimated even if the reinterviews and the original interviews were permitted to be dependent. This, however, requires a second reinterview survey, which, for practical reasons, may be difficult and certainly expensive to implement. Another example is that the early models almost always presuppose a simple sampling design, whereas, in practice, survey designs are usually much more complex.

iii. The experimental designs necessary for estimating the components in the early models are expensive to implement. When personal interviews are used, interpenetration of the interviewers' workloads can be very costly if the study area is large. This problem can be diminished if the population under study and the population of interviewers are stratified and the model is applied in each stratum, as suggested by Sukhatme, or if the populations are grouped as assumed in the Hansen et al. (1951) model. However, even these designs would be expensive for organizations such as Statistics Sweden where the interviewers are spread over the country and work alone in large areas. Another practical problem associated with interpenetration of interviewer assignments occurs in countries where the sampling units are individuals (and not housing units). Tracking respondents then becomes an important part of the interviewers' work. Since tracking respondents requires good knowledge of the local environment, interpenetrating could lead to increased nonresponse problems. In telephone interviews, the cost problems with interpenetration can almost be ignored, but the nonresponse problem cannot. Also, the costs of conducting reinterviews are considerable since large reinterview samples are needed for estimating the simple response variance component with an acceptable precision.

5. References

Abul-Ela, A. A., Greenberg, B. G., and Horvitz, D. G. (1967): A Multiproportions Randomized Response Model. Journal of the American Statistical Association, 62, pp. 990–1008.

Bailar, B. A. and Biemer, P. P. (1984): Some Methods for Evaluating Nonsampling Error in Household Censuses and Surveys. In Rao, P. S. R. S. and Sedransk, J. (eds.): W. G. Cochran's Impact on Statistics, pp. 253–274. John Wiley & Sons, New York.

Bailar, B. A. and Dalenius, T. (1969): Estimating the Response Variance Components of the U. S. Bureau of the Census' Survey Model. Sankhyā: Series B, pp. 341–360.

Casady, R. J. (1966): A Model for Analysis of Reinterview Data in the NCHS Health Interview Survey. Memo. U. S. Department of Health, Education, and Welfare.

Cochran, W. G. (1968): Errors of Measurement in Statistics. Technometrics, 10, pp. 637–666.

Collins, M. (1980): Interviewer Variability: A Review of the Problem. Methodological Working Paper No. 19. Social and Community Planning Research. London.

Deming, W. E. (1944): On Errors in Surveys. American Sociological Review, 9, pp. 359–369.

Des Raj (1968): Sampling Theory. McGraw-Hill, New York.

Fellegi, I. (1964): Response Variance and Its Estimation. Journal of the American Statistical Association, 59, pp. 1016–1041.

Forsman, G. (1987): Early Survey Models and Their Impact on Survey Quality Work. Proceedings of the U. S. Bureau of the Census' Third Annual Research Conference, Baltimore, MD, pp. 24–45.

Groves, R. M. and Magilavy, L. J. (1986): Measuring and Explaining Interviewer Effects in Centralized Telephone Surveys. Public Opinion Quarterly, 50, pp. 251–266.

Hansen, M. H., Hurwitz W. N., and Bershad, M. A. (1961): Measurement Errors in Censuses and Surveys. Bulletin of the International Statistical Institute, 38:2, pp. 359–374.

Hansen, M. H., Hurwitz, W. N., and Pritzker, L. (1964): The Estimation and Interpenetration of Gross Differences and the Simple Response Variance. In Rao, C. R. (ed.): Contributions to Statistics. Statistical Publishing Society, Calcutta.

Hansen, M. H., Hurwitz, W. N., Marks, E. S., and Mauldin, W. P. (1951): Response Errors in Surveys. Journal of the American Statistical Association, 46, pp. 147–190.

Hansen, M. H. and Tepping, B. J. (1969): Progress and Problems in Survey Methods and Theory Illustrated by the Work of the United States Bureau of the Census. In Johnson, N. L., and Smith, H. (eds.): New Developments in Survey Sampling. Wiley-Interscience, New York, pp. 1–26.

Jabine, T. B. and Tepping, B. J. (1973): Controlling the Quality of Occupation and Industry Data. Bulletin of the International Statistical Institute, 45, pp. 360–389.

Kish, L. (1962): Studies of Interviewer Variance for Attitudinal Variables. Journal of the American Statistical Association, 57, pp. 92–115.

Kish, L. and Lansing, J. B. (1954): Response Errors in Estimating the Value of Homes. Journal of the American Statistical Association, 49, pp. 520–538.

Krotki, K. P. and Hill, C. J. (1978): A Comparison of Correlated Response Variance Estimates Obtained in the 1961, 1971 and 1976 Censuses. Survey Methodology, 4, pp. 87–99.

Lyberg, I. (1986): Om evalveringsmetoder för Folk- och bostadsräkningen (FoB). Memo. Statistics Sweden. (In Swedish.)

Mahalanobis, P. C. (1946): Recent Experiments in Statistical Sampling in the Indian Statistical Institute. Journal of the Royal Statistical Society, 109, pp. 327–378.

Murthy, M. (1967): Sampling Theory and Methods. Statistical Publishing Society, Calcutta.

Palmer, G. L. (1943): Factors in the Variability of Response in Enumerative Studies. Journal of the American Statistical Association, 38, pp. 143–152.

Rice, S. A. (1929): Contagious Bias in the Interview. American Journal of Sociology, 35, pp. 420–423.

Sanchez-Crespo, J. L. (1973): Balance Between Sampling and Non-sampling Errors in Spanish Official Statistics. Bulletin of the International Statistical Institute, 45, pp. 329–367.

Sanchez-Crespo, J. L. (1981): Spanish Experience on the Estimation of Some Components of the Total Error. Statistical Commission and Economic Commission for Europe. Conference of European Statisticians, June 1–4, 1981.

Stock, J. S. and Hochstim, J. R. (1951): A Method of Measuring Interviewer Variability. Public Opinion Quarterly, 15, pp. 322–334.

Strecker, H., Wiegert, R., and Kafka, K. (1984): Practical Determination of a Response Variance on the Basis on Survey Models with Reenumerations. Jahrbücher für Nationalökonomie und Statistik, 199, pp. 1–31.

Sukhatme, P. V. (1954): Sampling Theory of Surveys With Applications. Rome: United Nations Food and Agriculture Organization. Published by: The Iowa State College Press and The Indian Society of Agricultural Research.

Sukhatme, P. V. and Seth, G. R. (1952): Nonsampling Errors in Surveys. Journal of the Indian Society of Agricultural Statistics, 4, pp. 5–41.

Swensson, B. (1969): Alternativa surveymodeller för dikotoma variabler. Report no. 19 of the Errors in Surveys research project. Institute of Statistics. University of Stockholm. (In Swedish.)

Tepping, B. J. and Boland, K. L. (1972): Response Variance in the Current Population Survey. Working Paper No. 36, Bureau of the Census. U. S. Government Printing Office, Washington, D. C.

U. S. Bureau of the Census (1978): The Current Population Survey. Design and Methodology. Technical Paper 40. U. S. Government Printing Office, Washington, D. C.

U. S. Bureau of the Census (1985): Evaluating Censuses of Population and Housing. Statistical Training Document, ISP-TR-5, Washington, D. C.

Warner, S. L. (1965): A Randomized Response Technique for Eliminating Evasive Answer Bias. Journal of the American Statistical Association, 60, pp. 63–69.

Zarkovich, S. S. (1966): Quality of Statistical Data. Rome: Food and Agriculture Organization of the United Nations.

Received May 1987
Revised January 1989