THE EFFECT OF MISMATCHING ON THE MEASUREMENT OF RESPONSE ERRORS*

JOHN NETER
University of Minnesota

E. SCOTT MAYNES
Bureau of the Census and University of Minnesota

R. RAMANATHAN
University of Minnesota

When response errors are studied by means of record checks, the possibility exists that matching errors are made in relating responses to the corresponding record data. Two simple models are developed for matching errors, and the implications of such matching errors on various measures of response errors are studied. The models are applied to the data from a record check study, and it is found that relatively small imperfections in the matching process can lead to substantial bias in estimating the relationship between response errors and "true" values.

1. INTRODUCTION

A MAJOR approach to the measurement of response errors is by record checks, or validation studies, in which survey responses are compared on a case-by-case basis with more-or-less accurate records. An important, though often unrecognized, obstacle to the usefulness and correct interpretation of record checks is the existence of matching errors. Matching errors arise when responses pertaining to one person (or family, organization, etc.) are incorrectly associated with and compared with record data pertaining to a different person. The impact of such matching errors on the measurement of response errors constitutes the core of this article.

An example will clarify the meaning of both record check and matching error. In the financial area one might undertake a record check study to measure the accuracy with which bank account balances are reported by their owners in sample surveys. The design of such a study—and the following description conforms roughly to a study actually being carried out by the Bureau of the Census—might be quite straightforward: (1) Draw a probability sample of bank accounts from bank records; (2) interview owners of sample accounts, asking them to provide complete information regarding each of the bank accounts they own; (3) compare information obtained in response to survey questions on an account-by-account basis with information in bank records.

For this design, matching errors (also called "mismatches") may arise in several ways: (1) the wrong person (not the owner of the sample account) is interviewed; (2) a bank clerk records the wrong bank balance (the balance of Account #53402 rather than of Account #53401); (3) analysts mistakenly match

* The authors happily acknowledge helpful comments by I. Richard Savage, Gad Nathan, Robert Ferber, and the referees.

the owner's report about his Account A in the sample bank with bank records pertaining to his Account B in the same bank. This list is by no means exhaustive.

In this paper the implications of matching errors are spelled out for two simplified models. The first model assumes, in terms of our example, that each account in the sample has the same probability of being matched correctly. In addition, if a mismatch occurs, the model assumes that each sample account has an equal probability of being mismatched with any other account in the parent population, regardless of the Account Number, the number of accounts possessed by this owner in the sample bank, the size of balance, the "commonness" of the owner's name (e.g., Smith), or other factors realistically related to the probability of mismatching.

The second model retains the assumption that mismatches occur according to a random mechanism, but permits mismatches to occur only within subsets of the parent population, and permits the probability of correct matching to differ among subsets. Because of these two features, the second model is more realistic and flexible. For example, if accounts owned by multiple-account holders (in the sample bank) are more likely to be mismatched, then "number of accounts owned in the sample bank" might be one of the criteria for defining population subsets in the second model. Subsets could be similarly defined to fit many other plausible hypotheses regarding sources of mismatching.

In actual record check studies, it is frequently found that it is impossible to match a sample account with any account in the parent population, thus giving rise to "nonmatches." Neither of our models deals with nonmatches. We have elected here to sacrifice realism in favor of simplicity.

As to the organization of this paper, Section 2 presents relevant background material on the conduct of record checks. Our two formal models are developed in Sections 3 and 4. Section 5 applies the models of Sections 3 and 4 to a particular example. Section 6 discusses other types of research where matching problems occur. Finally, our conclusions are summarized in Section 7.

2. RECORD CHECKS: ENCOUNTERS WITH MISMATCHING

An Example.—Consider Table 1, taken from a carefully conducted record check study by Horn [4]. For a sample of purportedly identical savings accounts, the table presents mean balances, as shown by bank records (Col. 2) and as reported by respondent-owners (Col. 3). Assuming perfect execution, accurate bank records, and that observed differences in means are statistically significant, the conclusion emerges inescapably from Col. 4 that respondents with large balances tended to underreport and respondents with small balances tended to overreport—in short, a regression-toward-the-mean effect.

But supposing that mismatches occurred—so that in some cases the respondent report and the bank report used to evaluate the accuracy of that particular respondent report do not refer to the same account. What would be the consequence? The answer is that mismatching could yield the same regression-toward-the-mean effect even in the case of zero response errors. Whether mismatching can explain the particular results of Table 1 or whether—by elimination—these results must be attributed to response errors will be discussed

TABLE 1. ACTUAL VS. REPORTED SAVINGS ACCOUNT BALANCES, OCTOBER, 1958 NETHERLANDS VALIDATION STUDY (GUILDERS)

(1) Groups of 100 Observations Ranked in Descending Order by Actual Balances	(2) Mean Actual Balance	(3) Mean Reported Balance	(4) Difference in Means: Reported Less Actual Balance
1	6110	5180	−930
2	4310	3820	−490
3	3530	3220	−310
4	3080	2610	−470
5	2730	2510	−220
6	2470	2280	−190
7	2130	2080	− 50
8	1790	1670	−120
9	1490	1410	− 80
10	1220	1170	− 50
11	1010	1160	+150
12	740	810	+ 70
13	480	550	+ 70
14	320	450	+130
15	180	270	+ 90
16	90	140	+ 50
17	10	170	+160

further in Section 5, when we use these data to illustrate formal results developed in Sections 3 and 4.

Before we turn to the formal analysis, however, we must deal with three important questions: (1) What factors give rise to mismatching? (2) What has been the frequency of mismatches in various matching studies? (3) How have the consequences of mismatching been dealt with analytically?

Sources of Mismatching.—Mismatching may occur through either (1) inadequacy of items available for matching, or (2) errors in the execution of the matching operation. We discuss each in turn.

Ideally, the set of items suitable for matching should define uniquely the thing being matched (a bank account, person, organization, etc.), be available in both sources of data, be measured or recorded accurately, and contain some items independent of the variable being studied.

There are several ways in which record checks can be designed so as to improve the probability of correct matches. The first way is to maximize the number and detail of items being used in matching. For instance, the number of "Robert Johnson's" in Minneapolis (telephone book count) is 224 (out of 334,000). Reduce the size of this subset by obtaining information about middle initials and you get 29 "Robert W. Johnson's." Finally, obtain (say) the name of wife and address, and the identification of a "Robert W. Johnson" in Minneapolis approaches uniqueness.

A second device to achieve uniqueness in matching—and the best if it is feasible—is by specifying items for matching which are in fact unique and provide a one-to-one mapping from one list to another, e.g., a bank account number in a particular bank or social security number (excepting the case where an individual maintains "aliases").

A third means of seeking uniqueness is by minimizing the size of lists in which persons are identified. In the bank account record check mentioned earlier, it would be better, *ceteris paribus,* to draw a sample from a small rural bank (with say, 15,000 accounts owned by people in small towns or rural areas) than to draw a sample from a New York City bank (with, say, 1–2 million accounts owned by people living mainly in a large metropolitan area).

Fourthly, it is desirable to locate the record check in a place (or list) which is as heterogeneous as possible with respect to the items used in matching. For example, it would be highly undesirable to match on surnames in Copenhagen with its abundance of Andersen's, Hansen's, etc.; by contrast, surnames may be more nearly unique if used for UN personnel in New York City.

Finally, greater accuracy of matching will be achieved if some information used for matching from the second source is obtained independently. For example, in the Census Bureau study mentioned earlier it was necessary to determine that the correct sample family was interviewed. Interviewers were supplied with the name of one savings account owner; names of other persons in the family were supplied by the person interviewed. When the names of other household members matched names on joint bank accounts, this was taken to be very strong evidence indeed that the correct family was interviewed.

An example of non-independence in matching is found in the same study. For families with multiple accounts, it was necessary to match individual accounts. One variable among several employed for this matching was the account balance: *cet. par.,* accounts were matched so as to minimize response errors.

Practical considerations have prevented many record check studies from employing optimal matching items. In some cases, a desire to protect the anonymity of respondents has forced some investigators to undertake matching without using the names of sample individuals [7]. Other studies have not asked respondents to supply their social security number or bank account numbers, either for fear of jeopardizing cooperation or because it was felt that the respondent could not or would not provide bank account numbers accurately.

As noted above, mismatching may also arise through errors in the execution of matching procedures. In general, mismatching from this source may be reduced by (1) minimizing the extent to which subjective judgments must be made, (2) replicating matches independently, and (3) utilizing consistency checks to detect errors due to carelessness.

Frequency of Mismatches.—Unfortunately, few data on frequency of mismatches are available [4, 5, 9]. A number of studies have provided information on nonmatches [6, 11, 12]. A nonmatch occurs when, using items available for matching, there appears to be no case in the parent population whose description conforms to a particular sample case. As a rough guide, nonmatched rates in typical studies have varied from a relatively small 1 per cent [12] to a rather large 36 per cent [11].

There are a number of factors which may account for this wide spread in the nonmatch rates encountered. The size and heterogeneity of the population studied, the quantity and quality of information available for matching, and the tightness of the rules employed for designating nonmatches, all bear on the nonmatch rate. In fact, it may well be that the tightness of the rules for nonmatches may cause the nonmatch rate and mismatch rate to vary inversely. Nevertheless, we believe that the nonmatch rate still is a useful indicator of the difficulty of matching.

Analytical Treatment of Mismatching.—In most matching studies, investigators have taken great pains to accomplish accurate matching. Due to differing underlying circumstances, their success in this has varied. What efforts were made analytically to take account of either detected or undetected mismatches? In some studies, particularly where the outcome of the matching procedures was obviously imprecise, analysts have designated various classes of matching, e.g., "positive matches," "probable matches," etc. When this procedure has been followed, it has been typical to confine most of the analysis to the "best" match class [11]. This has the possible undesirable effect of introducing bias, and also reduces sample size. On the other hand, it has the virtue of recognizing that mismatching may vitiate the statistical analysis unless corrective action is taken.

As far as residual, undetected mismatches go, this factor has not been explicitly dealt with in any of the studies with which we are familiar. Such mismatches are our primary concern.

3. MODEL 1: MATCHING ERRORS OCCURRING THROUGHOUT POPULATION

Nature of Model Studied.—We begin the study of the effects of matching errors on the measurement of response errors by considering a highly simplified model. As a vehicle for discussion, we shall use the example concerning the study of response errors in reporting of bank balances by household respondents. Suppose that the population consists of bank accounts $A_1, A_2, \cdots, A_N$. The balance of the jth account according to the bank records is denoted by Y_j $(j=1, 2, \cdots, N)$. These balances according to bank records are taken as the "true" values. Thus, the true mean balance per account in the population is:

$$\overline{Y} = \frac{1}{N} \sum_{j=1}^{N} Y_j \tag{1}$$

and the population variance of the account balances is:

$$\sigma_Y^2 = \frac{1}{N} \sum_{j=1}^{N} (Y_j - \overline{Y})^2 \tag{2}$$

We suppose now that the respondent for account A_j will report a balance W_j which is not subject to random errors. In other words, the simple model investigated here does not involve random response errors. Thus, the "true" response error R_j for the jth account is:

$$R_j = W_j - Y_j \tag{3}$$

For reasons mentioned in the previous section, matching errors may occur in the record check study, so that R_j may not be observed directly. Thus, the reported balance for the jth account may not be compared with the correct balance Y_j but with some other balance Y_k $(k \neq j)$. We therefore introduce a random variable Z_j for the jth account, which is defined as follows:

$$Z_j = \begin{matrix} Y_j \text{ with probability } p \\ Y_k \text{ with probability } q(k \neq j) \end{matrix} \tag{4}$$

Z_j represents the bank balance against which the reported balance W_j is compared. According to the simple model, the comparison is made against the correct balance with probability p, but may be made against any other bank balance in the population, with probability q for any specific alternate account. It follows therefore that:

$$p + (N - 1)q = 1 \tag{5}$$

This model has three important restrictive properties:

a. A match against some account must be made; thus, there is no provision for nonmatches in doubtful cases.

b. The probability of a correct match (p) is the same for all accounts, regardless of the amount of bank balance, number of accounts held in bank by same person, and so on.

c. The probability of a mismatch (q) is the same for all other bank accounts in the population.

The second and third limitations are relaxed in the following section. They are serious limitations. Often, the probability of a correct match will vary for different accounts. In fact, this probability may be correlated with the magnitude of the response error; for instance, a large response error may lead to additional search procedures which would tend to increase the probability of a correct match. Also, mismatching is probably more likely to occur within a small subset of the population accounts (for instance, within the accounts held by a family or by persons of the same name). We consider the case of equal probability of correct matches and possible mismatching throughout the population first because it is a simple case which provides considerable insights into the effects of matching errors, and because it serves as the foundation for the next model where the probability of correct matches may vary and matching errors are restricted within mutually exclusive subsets of the population.

The measured response error for the jth account is denoted by M_j, defined as follows:

$$M_j = W_j - Z_j \tag{6}$$

where M_j is a random variable since Z_j is a random variable. It follows from (4) that:

$$M_j = \begin{matrix} W_j - Y_j = R_j & \text{with probability } p \\ W_j - Y_k & \text{with probability } q(k \neq j) \end{matrix}$$

Thus, M_j provides the "true" response error only with probability p.

To summarize our basic notation in one location, we have:

$$\begin{aligned} Y &= \text{true bank balance} \\ W &= \text{reported bank balance} \\ R &= \text{true response error} \\ Z &= \text{matched bank balance} \\ M &= \text{measured response error} \end{aligned}$$

When an account is selected from the population at random, we denote the random variable corresponding to the measured response error as m, and similarly denote the random variables corresponding to the true balance and to the reported balance as y and w respectively.

A simple random sample of accounts with replacement is defined as one such that the w's are independent and the z's are independent. The condition that the z's are independent implies that the same account could be matched against several responses. If the survey matching procedures preclude duplicate matching, then the model may be appropriate only for larger populations where the probability of duplicate matching according to the model would be very small. On the other hand, if duplicate matching is possible—and this is the case in the matching studies with which we are familiar—the model permitting duplicate matching may be appropriate even for smaller populations.

Study of Mean Response Error.—In many investigations of response error, an important characteristic studied is the mean response error since it gives information on both the direction and magnitude of any bias present. We shall now show that matching errors do not effect the study of mean response errors with the model assumed.

We define the "true" mean response error as:

$$\overline{R} = \frac{1}{N}\sum_{j=1}^{N} R_j = \frac{1}{N}\sum_{j=1}^{N}(W_j - Y_j) = \overline{W} - \overline{Y} \tag{7}$$

where $\overline{W}$ and $\overline{Y}$ are defined as in (1).

We shall first obtain $E(m|j)$, the conditional mean of m, given that the jth population account was selected:

$$\begin{aligned} E(m \mid j) &= (W_j - Y_j)p + \sum_{k \neq j}(W_j - Y_k)q \\ &= (W_j - Y_j)(p - q) + Nq(\overline{W} - \overline{Y}) \end{aligned}$$

Taking expectations over all population accounts, we obtain:

$$\begin{aligned} E(m) = E\{E(m \mid j)\} &= \sum_{j=1}^{N}\left[(W_j - Y_j)(p - q) + Nq(\overline{W} - \overline{Y})\right]\frac{1}{N} \\ &= (p - q)(\overline{W} - \overline{Y}) + Nq(\overline{W} - \overline{Y}) \end{aligned}$$

Utilizing (5), we obtain:

$$E(m) = \overline{W} - \overline{Y} = \overline{R} \tag{8}$$

Thus, if a simple random sample of accounts is selected with replacement and the response errors measured, the mean measured response error of the sample,

$$\bar{m} = \left(\sum_{i=1}^{n} m_i\right) \Big/ n,$$

is an unbiased estimator of $\bar{R}$ even though matching errors are present.

Study of Response Error Variance.—Frequently, the variability of response errors is also of interest. We define the "true" response error variance as:

$$\sigma_R^2 = \frac{\sum_{j=1}^{N} (R_j - \bar{R})^2}{N} \tag{9}$$

To find the variance of m, we first obtain:

$$\begin{aligned} E\{(m - \bar{R})^2 \mid j\} &= p\{(W_j - Y_j) - (\bar{W} - \bar{Y})\}^2 \\ &\quad + q \sum_{k \neq j} \{(W_j - Y_k) - (\bar{W} - \bar{Y})\}^2 \end{aligned}$$

After some algebraic manipulation, we find:

$$\begin{aligned} E\{(m - \bar{R})^2 \mid j\} &= \{(W_j - Y_j) - (\bar{W} - \bar{Y})\}^2 + q \sum_{k=1}^{N} (Y_j - Y_k)^2 \\ &\quad + 2Nq(Y_j - \bar{Y})\{(W_j - Y_j) - (\bar{W} - \bar{Y})\} \end{aligned}$$

Taking the expectation over all population elements, we obtain:

$$\begin{aligned} \sigma_m^2 = E[E\{(m - \bar{R})^2 \mid j\}] &= \frac{1}{N} \sum_{j=1}^{N} \{(W_j - Y_j) - (\bar{W} - \bar{Y})\}^2 \\ &\quad + \frac{q}{N} \sum_{j} \sum_{k} (Y_j - Y_k)^2 + 2q \sum_{j} (Y_j - \bar{Y})\{(W_j - Y_j) - (\bar{W} - \bar{Y})\} \end{aligned}$$

The first term is σ_R^2 as defined in (9). The remaining two terms can be combined by algebraic manipulation and we obtain:

$$\sigma_m^2 = \sigma_R^2 + 2Nq\,\sigma_{WY} \tag{10}$$

where σ_{WY} is defined as follows:

$$\sigma_{WY} = \frac{\sum_{j=1}^{N} (W_j - \bar{W})(Y_j - \bar{Y})}{N}$$

It is clear from (10) that in the absence of matching errors ($q = 0$), $\sigma_m^2 = \sigma_R^2$. However, when matching errors are present, formula (10) indicates that the variance of the measured response errors is not equal to the variance of the true response errors except in the special case when $\sigma_{WY} = 0$. If the measurement technique utilized is effective to any extent, one would clearly expect the correlation between reported bank balances and true balances to be positive. In that case, we have:

$$\sigma_m^2 > \sigma_R^2$$

If, however, a substantial proportion of bank account owners respond that they have no account (so that $W_i = 0$ for these persons), it may happen that σ_{WY} is close to zero or even negative, even though there is a positive correlation between reported and true bank balances for persons acknowledging the ownership of an account. To be sure, if the proportion of persons not reporting the existence of an account is high enough to make σ_{WY} small positive or even negative under these conditions, the measurement technique is not effective. In the remainder of our paper, therefore, we shall concentrate the discussion on the case where σ_{WY} is positive.

If a simple random sample of accounts is selected with replacement, then the sample variance:

$$s_m^2 = \frac{\sum_{i=1}^{n} (m_i - \bar{m})^2}{n - 1} \tag{11}$$

is an unbiased estimator of σ_m^2, and would tend to overestimate the variance of the true response errors if σ_{WY} is positive.

It may be possible to obtain some indication of the magnitude of σ_R^2. Using (10), Appendix (A.3), and (5), we obtain:

$$\sigma_R^2 = \sigma_m^2 - 2Nq \frac{\sigma_{wz}}{1 - Nq} \tag{12}$$

Since s_m^2 as defined in (11) is an unbiased estimator of σ_m^2, and since:

$$s_{wz} = \frac{\sum_{i=1}^{n} (w_i - \bar{w})(z_i - \bar{z})}{n - 1} \tag{13}$$

is an unbiased estimator of σ_{wz} when using simple random sampling with replacement, an estimate of the relation between σ_R^2 and q can be established through (12). If σ_R^2 is quite insensitive to changes in q in the range where q is expected to fall, an indication of the magnitude of σ_R^2 may be obtained.[1]

Study of Relationship Between m and z.—Studies of response errors frequently consider a number of independent variables that might "explain" the response errors. For instance, the relationship between the magnitude of response errors and the amount of the bank balance might be of interest, or the relationship between response errors and income.

[1] An estimate of the maximum bias can also be obtained. The bias is maximum when $p=0$ or $q=1/(N-1)$. We thus have from (12):

$$|\sigma_m^2 - \sigma_R^2| \leq 2N|\sigma_{wz}|$$

and σ_{wz} can be estimated by s_{wz} as defined in (13).

We shall now consider the effect of matching errors on the covariance between m, the measured response error for a randomly selected account, and z, the matched bank balance. Our primary interest, to be sure, is in the covariance between the measured response error m and the true bank balance y, but the presence of matching errors precludes observations on the true values y.

In Appendix A, it is shown that:

$$\sigma_{mz} = \sigma_{RY} - qN\,\sigma_{WY} \tag{14}$$

Thus, in the usual case where σ_{WY} is positive and matching errors are present, we have:

$$\sigma_{mz} < \sigma_{RY}$$

so that the presence of matching errors would indicate a smaller covariance (algebraically) between response errors and true values than actually exists. It also follows, in view of (10) and (A.2), that matching errors lead to a smaller indicated correlation (algebraically) between response errors and true values than actually exists when σ_{WY} is positive.

It may be noted that if response errors are zero, the covariance between measured response errors (due to errors in matching) and the matched bank balances becomes:

$$\sigma_{mz} = -\,qN\,\sigma_Y^2 \tag{15}$$

since, then $\sigma_{RY}=0$ and $\sigma_{WY}=\sigma_Y^2$. Thus, mismatching would lead to a negative covariance between m and z in the absence of response errors.

If a linear regression between m and z were to be calculated, as a means of studying the nature of response errors, matching errors again usually would affect the analysis. It can be shown from (14) that the slope β_{mz} is given by:

$$\beta_{mz} = \beta_{RY} - qN\,\beta_{WY} \tag{16}$$

so that the presence of matching errors when the relation between W and Y is positive would lead to an understatement (algebraically) of the slope of the regression line of R on Y.

The constant α_{mz} in the linear regression equation is also affected by matching errors usually, since it follows that:

$$\alpha_{mz} = \alpha_{RY} + qN\,\beta_{WY}\overline{Y} \tag{17}$$

It may be noted, though, that if the "true" bank balances are expressed as deviations around the mean $\overline{Y}$, then the α term would not be affected by the presence of matching errors.

Figure 1 summarizes the situation where the reported and true bank balances are positively correlated, as are the "true" response errors and the correct bank balances. Matching errors lead to a smaller slope of the regression line of M on Z, and to a higher intercept on the M-axis.

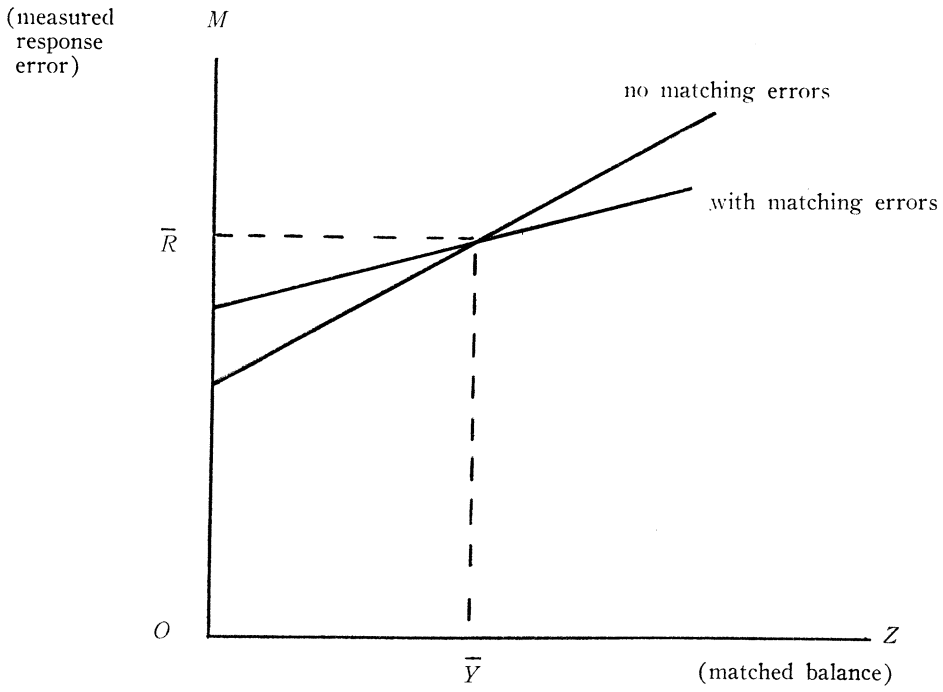

FIG. 1. Regression of measured response errors M on the matched bank balances Z, with and without matching errors, when both W and Y and R and Y are positively correlated.

If a simple random sample of bank accounts is selected with replacement, then an unbiased estimator of σ_{mz} is:

$$s_{mz} = \frac{\sum_{i=1}^{n} (m_i - \bar{m})(z_i - \bar{z})}{n - 1}$$

An indication of the magnitude of σ_{RY} may be obtained in a manner similar to that discussed for σ_R^2. It can be shown that:

$$\sigma_{RY} = \frac{\sigma_{mz} + \sigma_Y^2 Nq}{1 - Nq} \tag{18}$$

In record check studies, σ_Y^2 may be known. At any rate, it can be estimated from:

$$s_z^2 = \frac{\sum_{i=1}^{n} (z_i - \bar{z})^2}{n - 1}$$

which is an unbiased estimator of σ_Y^2 for simple random sampling with replacement. Also, s_{mz} is then an unbiased estimator of σ_{mz}, so that the relation between

σ_{RY} and q can be estimated, and the effect of varying q within the range where it is expected to fall can be studied.

Other Relationships.—At times, other relationships involving the response error may be of interest. The covariance between the reported balance and the corresponding matched bank balance is given by (A.3). This formula shows that if σ_{WY} is positive:

$$\sigma_{wz} < \sigma_{WY}$$

in the presence of matching errors.[2] Thus, the covariance between the reported and matched bank balances would be smaller in this case than the covariance between the reported and true bank balances. The same relation also holds for the correlation coefficients when σ_{WY} is positive.

The covariance between the measured response error m and the reported bank balance w is:

$$\sigma_{mw} = \sigma_{RW} + Nq\,\sigma_{WY} \tag{19}$$

Hence, when σ_{WY} is positive and matching errors are present, we have:

$$\sigma_{mw} > \sigma_{RW}$$

so that matching errors would lead to a larger covariance (algebraically) between measured response errors and reported bank balances than the covariance between the true response errors and reported bank balances.

Similarly, the covariance between the measured response error m and some other variable v furnished by the respondent (for instance, income) is given by:

$$\sigma_{mv} = \sigma_{RV} + Nq\,\sigma_{VY} \tag{20}$$

If the responses V_i are subject to responses errors, then the expression in (20) can be modified to reflect this.

If the measured response errors are to be compared with another bank account characteristic subject to the same matching error as the balance of the account (for instance, the length of time since the last transaction, according to the bank records), no new problem arises. The new variable, say X, simply replaces Y on the right side in (14), so that:[3]

$$\sigma_{mx} = \sigma_{RX} - Nq\,\sigma_{WX} \tag{21}$$

4. MODEL 2: MATCHING ERRORS RESTRICTED TO SUBSETS OF POPULATION

Nature of Model Studied.—In many cases it may not be realistic to assume that the probability of a correct match is the same for all elements in the population, nor that matching errors can occur throughout the population. Rather, such errors may be limited to subsets of the population, such as persons in a household, persons at the same address with the same name, or persons with the same name and age, and the probability of a correct match may vary from subset to subset. The subsets within which matching errors can occur depend on the specific matching techniques that are employed, and will vary from problem to problem.

[2] We do not consider the unlikely case that $p < 1/N$, when $(p-q)$ would be negative.

[3] The subscript x in σ_{mx} refers to a chance variable that is subject to matching errors.

The model considered in this section assumes that:

a. The population is divided into K mutually exclusive and exhaustive subsets.
b. Matching errors can occur only within a subset.
c. Within the ith subset, containing N_i elements, the probability of a correct match for any element is p_i, and the probability that any other particular element in the subset is used for the match is q_i. Thus we have:

$$p_i + (N_i - 1)q_i = 1 \tag{22}$$

given that an element from the ith subset is selected.

It is thus clear that the conditions within any subset correspond to those utilized in Section 3. Consequently, the derivations of results for the model in this section are an extension of those obtained earlier.

The limitations of the model discussed in the previous section still apply, namely that a match must be made, that the probability of a correct match is the same for all elements in a subset, and that mismatches against other elements are equally likely (but here only within the subset). In addition, Model 2 requires the subsets within which mismatches may occur to be mutually exclusive. This latter restriction often may not be a serious one, if the probability of a mismatch against elements outside the subset is very small compared to the probability of a mismatch within the subset.

To illustrate the nature of these subsets, we shall consider a record check study of bank balance reports. Here, for instance, mismatches may occur only within the group of accounts for persons with the same surname living at the same address. If, however, the probabilities of correct matches depend also on the bank balance, subsets meeting the requirements of the model discussed would have to be defined on three-dimensions: surname, address, and size of bank balance.

The notation to be used in this section is an extension of the earlier notation. There are K subsets in the population (households, for instance), the ith subset containing N_i elements (accounts, in our example). We let:

$$N = \sum_{i=1}^{K} N_i$$

The jth account in the ith household has a true balance Y_{ij} ($i=1, 2, \cdots, K$; $j=1, 2, \cdots, N_i$). Then the true mean balance per account in the i-th household is:

$$\overline{Y}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} Y_{ij} \tag{23}$$

and the true mean balance in the population is:

$$\overline{Y} = \frac{1}{N} \sum_i \sum_j Y_{ij} = \frac{1}{N} \sum_i N_i \overline{Y}_i \tag{24}$$

The "true" response error for the jth account in the ith household is defined as:

$$R_{ij} = W_{ij} - Y_{ij} \tag{25}$$

where W_{ij} is the reported balance (assumed to be fixed). The matched account balance is a random variable defined as follows:

$$Z_{ij} = \begin{matrix} Y_{ij} \text{ with probability } p_i \\ \\ Y_{ik} \text{ with probability } q_i \; (k \neq j) \end{matrix} \tag{26}$$

Then, the measured response error is:

$$M_{ij} = W_{ij} - Z_{ij} \tag{27}$$

which is a random variable because of (26).

Definitions of $\overline{R}_i$, $\overline{R}$, $\overline{W}_i$, $\overline{W}$ and so on parallel those given above.

Simple random sampling with replacement for this model is again defined so that the w's are independent and the z's are independent. The earlier discussion on possible limitations of the model because it permits duplicate matching applies here also.

Study of Mean Response Error.—As in Section 3, we find that matching errors do not bias the measurement of the mean response error. We have:

$$E(m) = E[E(m \mid i)]$$

where the term in brackets is the expectation of m given that the selected account fell in the ith household. But this expectation, from (8), is $\overline{R}_i$. Hence:

$$E(m) = \sum_i \overline{R}_i \left(\frac{N_i}{N}\right) = \overline{R} \tag{28}$$

since the probability that an account selected at random from the population falls into the ith household is N_i/N.

Study of Response Error Variance.—The effect of matching errors on the variability of the measured response errors is of the same type as with the model in Section 3. In Appendix B, we obtain:

$$\sigma_m^2 = \sigma_R^2 + \frac{2}{N} \sum_i N_i q_i \left[\sum_j (W_{ij} - \overline{W}_i)(Y_{ij} - \overline{Y}_i) \right] \tag{29}$$

Thus, if the reported and true bank balances are positively correlated within the subsets, then $\sigma_m{}^2$ overstates $\sigma_R{}^2$. [4]

As with the model in Section 3, the sample variance s_m^2 is an unbiased estimator of σ_m^2 when sampling is random with replacement as defined.

Study of Relationships Between Response Errors and Other Variables.—The effects of matching errors on the study of relationships between response errors and explanatory variables, when matching errors are restricted to subsets of the population, are similar to the effects noted for the model in Section 3.

[4] It may be noted that if the correlation in some subsets is positive, but negative in others, one cannot make a general statement about the direction of the bias.

Consequently, we shall simply present the covariances of primary interest, without further discussion of the nature of the effects or of special cases.

The covariance between m and y is:

$$\sigma_{my} = \sigma_{RY} + \frac{1}{N}\sum_i q_i N_i \left[\sum_j (Y_{ij} - \overline{Y}_i)^2 \right] \tag{30}$$

The covariance between m and z is:

$$\sigma_{mz} = \sigma_{RY} - \frac{1}{N}\sum_i q_i N_i \left[\sum_j (W_{ij} - \overline{W}_i)(Y_{ij} - \overline{Y}_i) \right] \tag{31}$$

The covariance between w and z is given by (B.6). Other covariances which may be of interest are:

$$\sigma_{mx} = \sigma_{RX} + \frac{1}{N}\sum_i q_i N_i \left[\sum_j (X_{ij} - \overline{X}_i)(Y_{ij} - \overline{Y}_i) \right] \tag{32}$$

$$\sigma_{mw} = \sigma_{RW} + \frac{1}{N}\sum_i q_i N_i \left[\sum_j (W_{ij} - \overline{W}_i)(Y_{ij} - \overline{Y}_i) \right] \tag{33}$$

5. APPLICATIONS OF THE MODELS

We shall now use the two models with the data obtained from the Horn record check study [4] in order to study further some implications of matching errors.

First, we examine the possibility that matching errors alone could account for the regression-toward-the-mean effect noted in Table 1. With model 1, the regression of measured response errors on the matched balance is, from (16) and (17):

$$E(m \mid z) = \alpha_{RY} + qN\beta_{WY}\overline{Y} + (\beta_{RY} - qN\beta_{WY})z$$

If there were no response errors, but only matching errors, $\alpha_{RY} = \beta_{RY} = 0$, $\beta_{WY} = 1$, and the regression equation would reduce to:

$$E(m \mid z) = qN\overline{Y} - qNz$$

Horn calculated for grouped data the unweighted regression of the measured response errors on the matched balances as:

$$\hat{m} = 202.6 - 0.178z$$

We can get estimates of q, assuming no response errors, from matching each of the two equation constants. Matching the slope terms, we have:

$$-qN = -0.178$$

or:

$$p = 1 - \left(\frac{N-1}{N}\right)0.178$$

Since N in this study was large, we obtain:

$$\hat{p} \cong 0.82$$

If we match the intercept terms, we obtain for large N:

$$\hat{p} \cong 1 - \frac{202.6}{\overline{Y}}$$

Horn estimated $\overline{Y}$ as 731, which gives $\hat{p} \cong 0.72$. Each of these estimates of p is made under the assumption that no response errors were present.

Before considering the reasonableness of the assumption of no response errors, we shall examine what the magnitude of p would be with model 2, assuming that no response errors were present. With this model, the slope of the regression between m and z is, using (31) and (A.2):

$$\beta_{mz} = \beta_{RY} - \frac{\sum_i N_i q_i \left[\sum_j (W_{ij} - \overline{W}_i)(Y_{ij} - \overline{Y}_i) \right]}{\sum_i \sum_j (Y_{ij} - \overline{Y})^2}$$

In the absence of response errors, this reduces to:

$$\beta_{mz} = \frac{-\sum_i N_i q_i \sum_j (Y_{ij} - \overline{Y}_i)^2}{\sum_i \sum_j (Y_{ij} - \overline{Y})^2}$$

If we now assume that the N_i's were reasonably large and roughly equal, and that the q_i's were approximately equal, we obtain for the case of no response errors:

$$p = 1 + \beta_{mz} \frac{\sum_i \sum_j (Y_{ij} - \overline{Y})^2}{\sum_i \sum_j (Y_{ij} - \overline{Y}_i)^2}$$

or, inserting the estimate for β_{mz}:

$$p = 1 - 0.178 \frac{\sum_i \sum_j (Y_{ij} - \overline{Y})^2}{\sum_i \sum_j (Y_{ij} - \overline{Y}_i)^2}$$

Since the ratio of the sums of squares cannot be less than 1, the estimate of p obtained with model 2 by matching the slope terms would not exceed the estimate of .82 obtained with model 1, given the restrictions assumed for model 2.

We have thus seen that if no response errors were present in the Horn study, the probability of a correct match would have had to be in the vicinity of .8 or less in order to account for the observed regression-toward-the-mean effect. Is this a reasonable probability for a correct match for this study? We believe not. The conductors of the Netherlands Validation Study took two major steps to minimize the possibility of mismatches: (1) The sample of the validation study was confined to persons owning, as far as bank records could disclose, but one account in the sample bank; this reduced the possibilities of accounts belonging to the same person being mismatched. (2) The investigators utilized considerable and accurate information to ascertain that the person interviewed was in fact the designated sample person: bank records provided the person's surname, given name, and address; government population registers gave information on the age of the depositor and the composition of his family.

In view of these steps to minimize matching errors, it is our judgment that the probability of a correct match for this study would be about .95 or higher. Thus, it appears to us highly unlikely that the negative slope of measured response errors on matched balances found by Horn is due to matching errors only, but rather reflects the behavior of response errors. It should be noted that a weighted regression estimate, taking into account the oversampling of large accounts by Horn, would have been preferable. We believe, however, that the conclusion reached about the presence of response errors would not be affected by this.

A second point which will be examined with the Horn data is the extent to which the slope of measured response errors on matched balances is biased because of matching errors. For model 1, the bias in the slope is, from (16):

$$\beta_{mz} - \beta_{RY} = - qN\beta_{WY}$$

This can be rewritten, assuming N is reasonably large, as:

$$\beta_{mz} - \beta_{RY} = \left(1 - \frac{1}{p}\right)(\beta_{mz} + 1)$$

We have the estimate of β_{mz} as -0.178. Hence, we can obtain the magnitude of the bias for different values of p. This has been done for a few selected values of p:

p	Absolute Bias	Relative Bias
1.00	0	0
.98	0.017	10 per cent
.95	0.043	24 per cent
.90	0.091	51 per cent

Thus, even if the probability of a correct match is as high as .98, the slope has a downward bias of about 10 per cent due to matching errors. Higher mismatching rates would involve still greater biases in the slope. It therefore appears that the consequences of even small mismatch rates can be considerable.

6. OTHER AREAS WHERE MATCHING PROBLEMS ARISE

In terms of applications, this paper has focused on studies of response errors, in particular record check studies. Matching problems arise in numerous other research studies.[5]

Some studies have involved the matching of multiple sets of established record systems. To study the relationship between certain economic-demographic factors and death rates, Hauser and Kitagawa [2], for example, merged death certificate information with economic-demographic information from the 1960 Census.

Other studies have required the matching of special research records with established record systems. One such study whose purpose was to ascertain mortality experience for victims of breast cancer involved the matching of records of X-ray and clinical examinations for a sample of women with death certificates for a part of the same sample [3].

Yet another application arises in the execution of reinterview and panel studies. The proper execution of such surveys requires that identical units be interviewed on each occasion. Failure to interview identical units will introduce mismatches of the type discussed in this paper.

Finally, matching problems are also encountered when analysts of survey data utilize information obtained from both the original sample list and from responses to survey questions. For example, in executing an alpha-segmental sample design, the New York Stock Exchange [8] obtained information from corporations and brokers on the number of stock issues owned by each sample individual. This information was then tabulated against information obtained from personal interviews—income, occupation and other characteristics.

7. CONCLUSIONS

We have considered two simple models for matching errors in order to study the effect of such errors on the analysis of response errors.

If matching errors can occur throughout the population in accordance with model 1, they will have the following effects:

a. Estimates of mean response errors will be unbiased.
b. The variance of response errors will be overstated when the correlation between the "true" and reported values is positive.
c. The slope of the regression of measured response errors on the matched values will be understated algebraically when the correlation between "true" and reported values is positive.

It was also noted that it may be possible to obtain an indication of the magnitude of the response error variance and of the correlation between response errors and "true" values if one can make a reasonable guess as to the range within which the probability q of a mismatch falls.

Model 2—where the probability of a correct match can vary and where mismatching occurs only within subsets—yields the same conclusions a-c as

[5] Shapiro and Densen [10] give a fairly complete, annotated list of applications.

model 1, provided that the correlations between "true" and reported values are in the same direction in each subset.

The models discussed need not refer only to the study of response errors. They may be applicable whenever some data are obtained from one source, other data from another source, and matching errors are possible. The models may also be extended in a number of directions. For instance, the restriction that the probability of a mismatched account is the same for all eligible accounts could be dropped, as could the assumption in model 2 that the probability of mismatching with an account outside the subset is zero. Similarly, the case when sampling is done within strata could be considered, as could the case when selection of population elements is done with unequal probabilities. In a related area, the effects of analyzing only data based on matches which are considered correct with high probability may be investigated.

APPENDIX

A. *Derivations for Model of Section 3*

Derivation of σ_z^2.—We first show that $Ez=\overline{Y}$. If the jth population element is selected:

$$E(z\,|\,j) = Y_jp + \sum_{k\neq j} Y_kq$$

in accordance with the definition of Z_j in (4). Taking expectations over all population elements, we obtain:

$$\begin{aligned} E(z) &= E\{E(z\,|\,j)\} = \sum_{j=1}^{N}\left[Y_jp + \sum_{k\neq j} Y_kq\right]\frac{1}{N} \\ &= \frac{1}{N}\sum_j [Y_j(p-q) + qN\overline{Y}] \\ &= (p-q)\overline{Y} + qN\overline{Y} \\ &= [p + q(N-1)]\overline{Y} \end{aligned}$$

so that:

$$E(z) = \overline{Y} \tag{A.1}$$

The last step follows in view of (5).

We next find the conditional expectation $E[(z-\overline{Y})^2|j]$. We have:

$$\begin{aligned} E[(z-\overline{Y})^2\,|\,j] &= (Y_j-\overline{Y})^2p + \sum_{k\neq j}(Y_k-\overline{Y})^2q \\ &= (Y_j-\overline{Y})^2(p-q) + \sum_{k=1}^{N}(Y_k-\overline{Y})^2q \end{aligned}$$

Taking expectations over all population elements, we obtain:

$$\sigma_z^2 = E\{E[(z-\overline{Y})^2 \mid j]\} = \sum_{j=1}^{N}\left[(Y_j-\overline{Y})^2(p-q) + \sum_{k=1}^{N}(Y_k-\overline{Y})^2 q\right]\frac{1}{N}$$
$$= (p-q)\sigma_Y^2 + qN\,\sigma_Y^2$$

Hence

$$\sigma_z^2 = \sigma_Y^2 \tag{A.2}$$

Derivation of σ_{wz}.—We first find the conditional expectation

$$E[(w-\overline{W})(z-\overline{Y}) \mid j],$$

given that the jth element in the population has been selected.

$$E[(w-\overline{W})(z-\overline{Y}) \mid j]$$
$$= (W_j-\overline{W})(Y_j-\overline{Y})p + \sum_{k\neq j}(W_j-\overline{W})(Y_k-\overline{Y})q$$
$$= (W_j-\overline{W})(Y_j-\overline{Y})(p-q) + \sum_{k=1}^{N}(W_j-\overline{W})(Y_k-\overline{Y})q$$

The second term on the right is zero, since $(W_j-\overline{W})$ is a constant over the summation. Proceeding then, we obtain:

$$\sigma_{wz} = E\{E[(w-\overline{W})(z-\overline{Y}) \mid j]\} = (p-q)\sum_{j=1}^{N}[(W_j-\overline{W})(Y_j-\overline{Y})]\frac{1}{N}$$
$$\sigma_{wz} = (p-q)\sigma_{WY} \tag{A.3}$$

Derivation of σ_{my}.—Since from (6):

$$m = w - z$$

we can write:

$$\sigma_{my} = \sigma_{w-z,y}$$

If a, b, c are chance variables, it is easy to show that:

$$\sigma_{a-b,c} = \sigma_{ac} - \sigma_{bc} \tag{A.4}$$

Consequently:

$$\sigma_{my} = \sigma_{wy} - \sigma_{zy}$$

Now:

$$\sigma_{wy} = \frac{\sum_{j}(W_j-\overline{W})(Y_j-\overline{Y})}{N} = \sigma_{WY}$$

and analogous to (A.3), we have:

$$\sigma_{zy} = (p-q)\sigma_Y^2$$

Hence:

$$\sigma_{my} = \sigma_{WY} - (p - q)\sigma_Y^2$$

Analogous to (A.4), it can be shown that:

$$\sigma_{RY} = \sigma_{WY} - \sigma_Y^2$$

We then obtain:

$$\sigma_{my} = \sigma_{RY} - (p - q - 1)\sigma_Y^2$$

Utilizing (5), this simplifies to:

$$\sigma_{my} = \sigma_{RY} + Nq\,\sigma_Y^2 \tag{A.5}$$

Derivation of σ_{mz}.—Using (A.4), we have:

$$\sigma_{mz} = \sigma_{wz} - \sigma_z^2$$

Using (A.2) and (A.3), we obtain:

$$\sigma_{mz} = (p - q)\sigma_{WY} - \sigma_Y^2$$

This expression can be simplified to:

$$\sigma_{mz} = \sigma_{RY} - qN\sigma_{WY} \tag{A.6}$$

B. *Derivations for Model of Section 4.*

Derivation of σ_m^2.—We first find Ez. If the selected population element is in the ith subset, we have from (A.1):

$$E(z \mid i) = \overline{Y}_i \tag{B.1}$$

Consequently, we obtain:

$$E(z) = E[E(z \mid i)] = \sum_i \overline{Y}_i\left(\frac{N_i}{N}\right) = \overline{Y} \tag{B.2}$$

Next we find σ_z^2. We utilize the following relationship [1, p. 65, formula 6.3]:

$$\sigma_z^2 = E\sigma_{z|i}^2 + \sigma_{E(z|i)}^2 \tag{B.3}$$

Now, $\sigma_{z|i}^2$ is given by (A.2). Consequently, we have:

$$E\sigma_{z|i}^2 = \sum_i\left[\sum_j \frac{(Y_{ij} - \overline{Y}_i)^2}{N_i}\right]\left(\frac{N_i}{N}\right) = \sum_i\sum_j \frac{(Y_{ij} - \overline{Y}_i)^2}{N}$$

Further:

$$\sigma_{E(z|i)}^2 = E[E(z \mid i)]^2 - [E\{E(z \mid i)\}]^2$$

Using (A.1) and (B.2), we obtain:

$$\sigma^2_{E(z|i)} = \sum_i \overline{Y}_i^2\left(\frac{N_i}{N}\right) - \overline{Y}^2$$

$$= \frac{1}{N}\sum_i N_i(\overline{Y}_i - \overline{Y})^2$$

Thus we have:

$$\sigma^2_z = \frac{1}{N}\sum_i\sum_j (Y_{ij} - \overline{Y}_i)^2 + \frac{1}{N}\sum_i N_i(\overline{Y}_i - \overline{Y})^2$$

or [1, p. 129, formula 6.1]:

$$\sigma^2_z = \sigma^2_Y \tag{B.4}$$

To find σ^2_m, we utilize:

$$\sigma^2_m = \sigma^2_{w-z} = \sigma^2_w + \sigma^2_z - 2\sigma_{wz}$$

Now:

$$\sigma^2_w = \sigma^2_W = \frac{1}{N}\sum_i\sum_j (W_{ij} - \overline{W})^2$$

Using (B.4) and (B.6), we have:

$$\sigma^2_m = \sigma^2_R + \frac{2}{N}\sum_i N_i q_i\left[\sum_j (W_{ij} - \overline{W}_i)(Y_{ij} - \overline{Y}_i)\right] \tag{B.5}$$

since:

$$\sigma^2_R = \sigma^2_Y + \sigma^2_W - 2\sigma_{YW}$$

Derivation of σ_{wz}.—We know [1, p. 66, formula 6.4] that:

$$\sigma_{wz} = E\,\sigma_{wz|i} + \sigma_{E(w|i)E(z|i)}$$

Now, $\sigma_{wz|i}$ is given by (A.3), so that:

$$E\,\sigma_{wz|i} = \sum_i\left[(p_i - q_i)\frac{1}{N_i}\sum_j (W_{ij} - \overline{W}_i)(Y_{ij} - \overline{Y}_i)\right]\left(\frac{N_i}{N}\right)$$

Further, from (B.1) and (B.2) and the corresponding results for w, we have:

$$\sigma_{E(w|i)E(z|i)} = \sum_i (\overline{W}_i - \overline{W})(\overline{Y}_i - \overline{Y})\left(\frac{N_i}{N}\right)$$

Hence, we obtain:

$$\sigma_{wz} = \sum_i \left[(p_i - q_i) \frac{1}{N_i} \sum_j (W_{ij} - \overline{W}_i)(Y_{ij} - \overline{Y}_i) \right] \left(\frac{N_i}{N} \right)$$
$$+ \frac{1}{N} \sum_i N_i(\overline{W}_i - \overline{W})(\overline{Y}_i - \overline{Y})$$

Using (22), we have:

$$\sigma_{wz} = \sigma_{WY} - \sum_i N_i q_i \left[\frac{1}{N} \sum_j (W_{ij} - \overline{W}_i)(Y_{ij} - \overline{Y}_i) \right] \qquad \text{(B.6)}$$

REFERENCES

[1] Hansen, Morris H., Hurwitz, William N., and Madow, William G., *Sample Survey Methods and Theory,* Volume II. New York: John Wiley, 1953.

[2] Hauser, Philip M., and Kitagawa, Evelyn, "Social and Economic Mortality Differentials in the U. S., 1960: Outline of a Research Project," *Proceedings of the Social Statistics Section,* American Statistical Association, 1960, 116–21.

[3] Health Insurance Plan of Greater New York, *Annual Statistical Report,* 1962.

[4] Horn, W., "Reliability Survey, A Survey on the Reliability of Responses to an Interview Survey," Reprint of an article appearing in *Het PTT-bedrijf,* 10 (1960).

[5] Horn, W., "Non-Response in an Interview Survey," Reprint of an article appearing in *Het PTT-bedrijf,* 12 (1963).

[6] Kahn, Robert L., *A Comparison of Two Methods of Collecting Data for Social Research: The Fixed Alternative Questionnaire and the Open-Ended Interview.* Ann Arbor: University of Michigan, Ph.D. Dissertation, 1952.

[7] Lansing, John B., Ginsburg, Gerald P., and Braaten, Kaisa, *An Investigation of Response Error,* Studies in Consumer Savings, No. 2. Urbana, Illinois: Bureau of Economic and Business Research, 1961.

[8] New York Stock Exchange, Department of Research and Statistics, *Methodology and Sample Design of 1962 Census of Shareowners.* New York: New York Exchange, 1962.

[9] Phillips, William Jr., and Bahn, Anita K., "Experience with Computer Matching of Names," paper presented at the September, 1963 Meetings of the American Statistical Association, Cleveland.

[10] Shapiro, Sam, and Densen, Paul M., "Research Needs for Record Matching," paper presented at the September, 1963 Meetings of the American Statistical Association, Cleveland.

[11] Sirken, Monroe G., Maynes, E. Scott, and Frechtling, John A., "The Survey of Consumer Finances and the Census Quality Check," in National Bureau of Economic Research, *An Appraisal of the 1950 Census Income Data,* Studies in Income and Wealth, Volume 23. Princeton: Princeton University Press, 1958, pp. 127–68.

[12] U. S. National Health Survey, *Reporting of Hospitalization in the Health Interview, A Methodological Study of Several Factors Affecting the Reporting of Hospital Episodes.* Washington: U. S. Department of Health, Education, and Welfare, Public Health Service, 1961, Publication No. 584-D4.