THE EFFECT OF MISMATCHING ON THE MEASUREMENT OF RESPONSE ERRORS*

John Neter, University of Minnesota,
E. Scott Maynes, U. S. Bureau of the Census and University of Minnesota
and R. Ramanathan, University of Minnesota

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 erors 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) Do 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 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 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 "uncommonness" 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 mismatching 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.

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 later.

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

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

TABLE 1

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

(1)	(2)	(3)	(4)
Groups of 100 Observations Ranked in Descending Order by Actual Balances	Mean Actual Balance	Mean Reported Balance	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

execution of the matching operation. We discuss each in turn.

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

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).

Finally, 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, say, for UN personnel in New York City.

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. A number of studies have provided information on nonmatches, which may be considered a proxy variable for mismatches in the sense of indicating the difficulty of matching. 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.

Past matching studies have varied considerably with respect to reported rates of mismatch or nonmatch. As Table 2 shows, reported mismatch or nonmatch rates vary from an inconsequential 0.4 percent in the Horn study, to a rather large 35.5 percent in the Sirken study.[1] It should be noted that the reported rate of mismatch or nonmatch may differ in either direction from the actual rate: characteristics of actually identical persons may be recorded erroneously in either of the two sets of records on which matches are based; alternatively, persons with apparently identical characteristics (e.g., the same name, same age, same sex, etc.) may in fact not be the same individuals.

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.

[1] We have neglected the 80.1 percent mismatch rate in the Phillips study (line 5) since the computer match was viewed as but one of two stages of matching.

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 an example concerning the study of response errors in reporting of bank balances by household respondents. Suppose that the population consists of bank accounts $A_1, A_2, \ldots A_N$. The balance of the j-th account according to the bank records is denoted by Y_j $(j = 1, 2, \ldots N)$. These balances according to bank records are taken as the "true" values. Thus, the true mean balance per account in the population is:

$$\bar{Y} = \frac{1}{N} \sum_{j=1}^{N} Y_j \qquad (1)$$

and the population variance of the account balances is:

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

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 j-th account is:

$$R_j = W_j - Y_j \qquad (3)$$

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 j-th account may not be compared with the correct balance Y_j but with some other balance Y_j $(k \neq j)$. We therefore introduce a random variable Z_j for the j-th account, which is defined as follows:

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

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:

$$p + (N - 1)q = 1 \qquad (5)$$

This model has two important restrictive properties:

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

b. If any mismatching occurs, any other bank balance is equally likely to be the mismatched balance.

TABLE 2

FREQUENCY OF NONMATCHES OR MISMATCHES: SELECTED MATCHING STUDIES

Reported Rate of Nonmatches or Mismatches	Sample Description	Variables Being Matched: First Source	Variables Being Matched: with Second Source	Information Used in Matching	Reference
1. 0.4% mismatch[a]	3321 savings acct. owners in 3 metrop. areas in Netherlands	Savings accounts and owners as reported in survey interviews	Accounts and owners as shown in bank records	Name, address, age family composition	Horn [4,5]
2. Two stages:					
a. 1% nonmatch	1491 persons in NC and NE U. S. who had been hospitalized	Persons as identified by interviewers	Persons as shown by hospital records	Name, address, age, sex, race	U. S. NHS [12]
b. 3.6% nonmatch	"	Hospitalization episodes as reported in interview	Hosp. episodes as shown in hospital records	Subjective matching by two persons	
3. 21% nonmatch	206 workers in single plant	Persons, as identified from answers to pencil-and-paper tests	Persons, as shown by plant records	Age, sex, section of plant where works, shift, etc.	Kahn [6]
4. 35.5% nonmatch	National sample of 1500 families	Families and related individuals identified by one survey organization	Families and related individuals identified by a second survey organization	Head's sex, age, occupation, veteran status, family size, no. of children, no. born in 1949-50 (special weight given to unusual characteristics)	Sirken [11]
5. Two Classes of computer matches[b]	22,869 psychiatric case records, incl. in some cases more than 1 record per person	Case records of particular persons (1961)	Any other case records among the 22,869 pertaining to the same person	Sequential comparisons on soundex code, surname, first name, address, birth year range, soc. security no., maiden name, sex-race, birth month and day, birth year	Phillips [9]
a. 3.7% mismatch	627 "positive matches" as determined by computer	"	"		
b. 80.1% mismatch	1,011 "possible matches"	"	"		

[a]Actually, matching with respect to family composition and age was carried out subsequent to the initial matching on name and address. The family composition-age check disclosed 60 actual mismatches (1.8% of the sample) which had been incorrectly accepted as matches. Of these, 47 cases were deleted from the sample before analysis; the remaining 0.4% were detected after analysis.

[b]The object was to eliminate duplicate records (finally determined to be 805) from the 22,869 records. This was achieved in two stages, first by a computer check which identified "positive matches" and "possible matches" and second by a careful clerical check which produced the mismatch rates shown in the table. Some of the records were incomplete with respect to the items used for checking.

The second limitation is relaxed in the following section. It is a serious limitation since 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 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 matching errors are restricted within mutually exclusive subsets of the population.

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

$$M_j = W_j - Z_j \qquad (6)$$

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

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

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

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

Y = true bank balance
W = reported bank balance
R = true response error
Z = matched bank balance
M = measured response error

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

A simple random sample of accounts with replacement is defined as one such that the $\underline{w}$'s are independent and the $\underline{z}$'s are independent. The condition that the $\underline{z}$'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.

Results -- We shall now state the major results, without giving any of the derivations:

1. With the model assumed, matching errors do not affect the study of mean response errors. It can be shown that:

$$E(m) = \bar{R} \qquad (7)$$

where $\bar{R}$ is the mean of the "true" response errors for the population. Hence, if a simple random sample of accounts is selected with replacement and the response errors measured, the mean measured response error of the sample is an unbiased estimator of $\bar{R}$ even though matching errors are present.

2. It also follows for this model that:

$$\sigma_m^2 = \sigma_R^2 + 2Nq\,\sigma_{WY} \qquad (8)$$

where σ_R^2 is the variance of the true response errors for the population and σ_{WY} is the covariance between the reported bank balances and the corresponding true bank balances in the population. Thus, the variance of the measured response errors is in general different from the variance of the true response errors. For instance, if σ_{WY} is positive, σ_m^2 would then exceed σ_R^2.

3. If a linear regression between the measured response error m and the matched bank balance z is calculated, then it can be shown that for this model, we have:

$$\beta_{mz} = \beta_{RY} - qN\beta_{WY} \qquad (9)$$

and:

$$\alpha_{mz} = \alpha_{RY} + qN\beta_{WY}\bar{Y} \qquad (10)$$

Thus, if the correlation between W and Y is positive:

$$\beta_{mz} < \beta_{RY}$$

and, assuming $\bar{Y}$ is also positive:

$$\alpha_{mz} > \alpha_{RY}$$

In other words, for the typical case where σ_{WY} is positive and $\bar{Y}$ is positive, the regression between the measured response error $\underline{m}$ and the matched bank balance $\underline{z}$ involves a smaller slope and larger intercept than the regression of the true response error R on the true bank balance Y.

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 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. The subsets within which matching errors can occur depend on the specific matching techniques that are employed, and will vary from problem to problem.

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 i-th 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:

$$p_i + (N_i - 1)q_i = 1 \qquad (11)$$

given that an element from the i-th 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 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 be met approximately, as when 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 mismatching probabilities 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.

Results -- Again, we simply present results without showing derivations:

1. As in the case of Model 1, the expectation of the measured response error $\underline{m}$ is $\bar{R}$.

2. Model 2 yields the same conclusions concerning the variance of $\underline{m}$ and the regression of $\underline{m}$ on $\underline{z}$ as Model 1, provided that the correlations between true and reported values are in the same direction in each subset.

5. APPLICATION OF THE MODELS

We shall now apply the earlier results to the data obtained from the Horn record check study [4] in order to 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 (9) and (10):

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

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:

$$E(m|z) = qN\bar{Y} - qN\,z$$

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

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

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:

$$\hat{p} \simeq 0.82$$

Thus, 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 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 a variety of steps to minimize the possibility of mismatches.

Their matching procedures were so thorough that 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.

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 Stock 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.