THE MATCHING ERROR STUDY FOR THE 1988 DRESS REHEARSAL

Mary H. Mulry and Aref Dajani. U. S. Bureau of the Census*
Washington, DC
Paul P. Biemer, University of New Mexico

1. INTRODUCTION

The U. S. Bureau of the Census conducted a Post Enumeration Survey (PES) to evaluate the census coverage error in the 1988 Dress Rehearsal. Currently, the Census Bureau plans to use the same methodology to evaluate the 1990 Decennial Census. This paper discusses the Matching Error Study, an evaluation of the processing of the data from the PES following the Dress Rehearsal. The project is one of several in the Evaluation of Census Coverage Estimates. Similar studies are under consideration for the PES estimates of census coverage error for 1990.

The PES was really two samples. The PES was composed of the E sample, which is a sample of census enumerations, and the P sample, which is a sample of the population independent of the census. The E sample and the P sample were overlapping, which means that the same blocks were selected for each sample. The E sample measured the erroneous census enumerations in the census and the P sample measured census omissions. Together they were used in dual system estimation to produce an estimate of the census coverage error.

The definition of the dual system estimator of the population size used by the Census Bureau first requires the following definitions:

CEN = the size of the original enumeration

II_1 = the number of persons imputed

II_2 = the weighted number of census enumerations with insufficient information for matching

EE = the estimate of the number of erroneous enumerations in the original enumeration

$C = CEN - II_1 - II_2 - EE$

N_p = the estimate of the total population from the P sample

M = the estimate of the number of people in the Census and the P sample

The dual system estimator then is given by

$$N = CN_p/M.$$

Dual system estimation assumes that the P-sample respondents can be linked, or matched, correctly to their census enumerations. Also, there is the assumption that census enumerations in the E sample can be properly identified as correct or erroneous.

The goal of the Matching Error Study was to estimate the bias in M and EE due to processing error in the assignment of enumeration statuses. These were errors that occurred even when the people were real and the Census Day address was reported correctly. Therefore, the error measured did not encompass response errors that arose in the data collection. Other types of errors that may result in an inaccurate assignment of a respondent's census enumeration status or match status were evaluated in other evaluation studies. The results are not yet available.

We were concerned about processing error in M and EE because biases in these estimates may cause a bias in the dual system estimator. Processing error is caused by two sources, procedures and personnel. Procedure error arises from error in the procedures designed for the matching operation. Personnel error is caused by error in the application of the procedures through mistakes or otherwise departing from the procedures.

Of course, in order to estimate the bias in M and EE, the true enumeration status must be known for a sample of the P-sample and E-sample cases. Since we cannot know the truth for every case, we must settle for obtaining the "best possible" status for a sample of cases. We then use the "best possible" status to estimate the bias in M and EE.

The Matching Error Study had two phases, one to measure error from each source of processing error. Phase I estimated personnel error and its effect on the PES estimates. Phase II estimated procedure error and its effect on the PES estimates.

Previous work on the effect of the failure to accurately match persons from the P sample to the census has been done by Seltzer and Adlakta (1974) and Scheuren and Oh (1985). Wolter (1983) stated that suspected matching errors in the 1980 Post Enumeration Program were a part of the reason not to adjust the 1980 Census. Biemer (1988) proposed a model for investigating the effect of matching error on the estimator of census coverage error. A pilot study for the Matching Error Study was conducted with data from the PES for the 1986 Test Census (Corby and Mulry, 1988).

Section 2 of this paper describes the study design, including the sample design. Section 3 reports the results of the study. Conclusions are given in Section 4.

2. STUDY DESIGN

2.1 Production Matching

After the 1988 PES interviewing, the matching operation determined whether the P-sample respondents were enumerated in the census and whether the E-sample cases were correct enumerations. Some cases were sent for a follow-up interview to collect additional

information. At the end of the operation, each P-sample case was designated as matching an enumeration, not matching an enumeration, or unresolved. The E-sample cases were designated as correctly enumerated, erroneously enumerated or unresolved. The following types of enumerations were considered erroneous; (1) people who died before Census Day, (2) people whoe were born after Census Day, (3) enumerations that do not refer to real people, (4) people duplicated, (5) people enumerated outside the search area where the matching operation looked for their enumeration. The search area for a case included the block for its address and the ring of adjacent blocks. A probablity of a correct enumeration status was imputed for each unresolved P-sample and E-sample case.

Matching occurred in two major phases, before field follow-up and after field follow-up. The first step was a computer matching operation. The computer used the Felligi-Sunter (1969) algorithm to match the P sample and the census. The computer designated matches, possible matches, P sample nonmatches, and E sample nonmatches. Next was a two-level clerical operation. The first level clerks were given rules for designating matches. The second level clerks, called the Special Matching Group (SMGs), had more flexibility and were able to use their judgment in designating matches.

For quality control, a second SMG clerk "independently" matched each block. "Independently" means that the matchers did not have access to the match codes assigned by the first-level clerk and the first SMG clerk. The results from the first-level clerk and the first SMG clerk were compared with the results from the second SMG clerk. When there were different match codes assigned, the differences were adjudicated by the PES Technicians who substituted a reconciled code. PES Techncians performed adjudication for the before follow-up matching. A computerized quality control system kept track of the codes assigned during the various steps of the matching operation.

When a P-Sample or E-Sample case could not be resolved, it was sent back to the field for follow-up. After the follow-up, the SMG clerks used the new information and attempted to resolve the case. For quality control, two SMG clerks matching the after follow-up cases. When the two SMG clerks disagreed, a PES technician adjudicated the case and substituted a final match code.

2.2 Phase I Design

Both Phase I and Phase II were conducted at the conclusion of the PES matching operation. Phase I was designed to achieve the "best match classification". Then we used the best match classification to compute error rates for the production. Phase I consisted of an "independent" rematch of a subsample of 100 of the PES blocks by special matching group personnel working in teams of two after they completed the production matching. "Independent" meant that the matchers did not have access to the production match codes. The underlying assumption was that personnel with more training, working in teams of two, and without the time pressure of production make fewer mistakes in classifying enumeration statuses although they have the same materials and information available as the production matchers. An expert matcher then reviewed both sets of match codes assigned to the difficult cases and made a decision on the match codes.

To define the "difficult" cases, we view the P sample (the E sample can be considered analogously) as being divided into two classes: Cases whose match statuses are obvious and unequivocal, and cases which are more difficult. Ideally, since almost all matching errors will occur in the class of "difficult" cases, expert matchers should only need to consider these cases. An accepted method for separating cases into these two classes is the match-rematch method. There are only three basic outcomes of a match attempt: (a) a match is found (M), (b) a match is not found (NM), and (c) the match status is unresolved (U). Actually, each basic outcome has several subcategories. However for illustrative purposes, nine possible basic outcomes are possible as indicated in Table 2.1 .

Table 2.1 Nine Basic Outcomes

		Rematch		
		M	NM	U
	M	(M,M)	(M,NM)	(M,U)
Match	NM	(NM,M)	(NM,NM)	(NM,U)
	U	(U,M)	(U,NM)	(U,U)

Table 2.1 can be used to categorize cases as either "not difficult" and "difficult." The scheme used in the MES is as:

> Cases classified as (M,M), (NM,NM) and (U,U) are categorized as "not difficult." An expert matcher would not be concerned with these.
>
> Cases classified in the off-diagonal cells of the table would be categorized as "difficult" and sent to the expert matcher for matching.
>
> If a case is labeled "not difficult," it is assumed to be classified correctly by the original match. If it is labeled as "difficult," it is assumed that the expert matcher classifies the case correctly. Both categories of cases, therefore, arrive at a "best match classification".

In the actual implementation, the "difficult" category was expanded to include cases where there was disagreement between the match and

rematch subcategories of the basic outcomes, M, NM, and U.

2.3 Phase II Design

In Phase II, a team of professionals with expertise in census operations and PES operations conducted a dependent rematch of a sample of blocks, with the option of field work when clarification was needed. "Dependent" meant that the matchers had access to the match codes assigned by the PES production matching. The professionals were not restricted by the procedures, but were instructed to use their best judgment and their knowledge in assigning match codes. In this phase, a team of specialists reviewed discrepancies between the production final codes and the rematch codes. The sample for Phase II consisted of the same 78 blocks selected for Phase I for the St. Louis and East Central Missouri sites.

2.4 Sample Size

The total sample size required for the 88 MES is 100 blocks or 7,200 persons assuming an average block size of 72 persons. The sample was allocated to the nine strata as shown in Table 2.2.

Table 2.2 Sample Sizes for the Matching Error Study

Stratum	Sample Size (Blocks)	Sample Size (Number of Persons)
St. Louis City	38	3,647
Black Renters	16	1,527
Black Owners	11	1,200
Non-Blacks	11	920
East Central MO	40	2,423
TAR	8	668
Prelist	16	976
UP/LV	16	779
East WA	22	1,627
TAR	4	252
Prelist	7	434
UP/LV	11	941
Total	100	7,697

3. RESULTS

3.1 Limitations of the Study

In order to interpret the results of the 1988 MES appropriately, some discussion of the limitations of the study is necessary. There are several ways in which the present study differs from the study which will be conducted following a 1990 PES; therefore, inferences regarding the 1990 experience from the 1988 data are limited. These are:

1. For the 1990 Census, both the PES and the MES will be many times larger than the 1988 Dress Rehearsal PES and 1988 MES. This "scale-up" could have an effect on a number of facets of the PES and MES operations which in turn will have an effect on the quality of the match and match evaluation results. Among these factors are: personnel, supervisory controls and quality controls. Therefore, the level of error measured in our study should not be interpreted as being representative of the level of error expected in 1990.
2. The procedures used in the 1988 MES will undergo some modifications for 1990. These changes could have an effect on estimates of matching error for 1990.
3. In this study, it was necessary to use only one reconciler (a match procedures expert on the Undercount Research Staff) to determine the correct match code in the case of a discrepancy between match codes and rematch codes in the MES. Thus, to the extent that the adjudicator imparts a systematic error in the final "true" match codes, the estimates of matching bias will be biased estimates. In 1990, a number of reconcilers will be employed.

3.2 Phase I Nonmovers

Tables 3.1 through 3.12 display the results of Phase I for nonmovers. Tables 3.1 through 3.3 have unreconciled P-sample results for St. Louis, East Central Missouri, and Washington State, respectively. In these tables a case is classified as enumerated, not enumerated, or unresolved. The unresolved status means that the clerks could not determine whether the person was enumerated. The E-sample results are shown in Tables 3.4 through 3.6 for the sites in the same order. In these tables, cases are classified as correctly enumerated, erroneously enumerated, or unresolved. The unresolved status means that the clerks could not determine whether the enumeration was correct.

For the unreconciled data, the gross difference rate provides a measure of the disagreement between the production and the rematchers. The gross difference rate is the sum of the off-diagonal cells divided by the total number of cases considered. For the P-sample, the gross difference rate is 0.019 for St. Louis, 0.002 for East Central Missouri, and 0.024 for Washington State. A model-based estimate of the standard error for the gross difference rate (Hansen, Hurwitz and Pritzker, 1964) is 0.002 for St. Louis, 0.001 for East Central Missouri and .004 for Washington State. Model-based estimates of the standard errors are more conservative than designed-based estimates. Design-based estimates are not available at this time. For the E-sample,

the gross difference rate is 0.038 for St. Louis, 0.028 for East Central Missouri, and 0.015 for Washington State. The estimated standard error is .003 for all three sites.

Tables 3.7 through 3.12 display the reconciled data. The P-sample results are shown in Tables 3.7 through 3.9 for St. Louis, East Central Missouri, and Washington State, respectively. The E-sample results are displayed in Tables 3.10 through 3.12 for the three sites in the same order.

For the reconciled P-sample data, the net difference in the observed match rate is the primary estimate of interest because it effects the bias in the dual system estimator. The match rate is calculated by dividing the matches by the sum of the matches and the nonmatches. Cases which are classified as unresolved by the original production and the rematch are not included in this calculation because these cases will be assigned imputed probabilities of matching. With the St. Louis data from the P sample, the match rate is 85.2 percent from the production and 85.2 the rematch. The study measured no net difference in the match rate with an estimated standard error of 0.2 percent. With the East Central Missouri data for the P sample, the match rate is 93.8 percent from the production and 93.9 percent from the rematch. The net difference in the match rate is -0.1 percent with an estimated standard error of 0.1 percent. For Washington State, the match rate is 93.6 percent from the production and 92.9 percent from the rematch. The net difference is 0.7 percent with an estimated standard error of 0.4 percent.

For the reconciled E-sample results, the estimate of primary interest is the percentage of erroneous enumerations because it effects the bias in the Bureau's dual system estimator. In St. Louis, the rate of erroneous enumerations is 2.4 percent in the production and 2.3 percent in the rematch. The net difference is 0.1 percent with an estimated standard error of 0.3 percent. In East Central Missouri, the rate of erroneous enumerations is 2.1 percent in the production and 2.2 percent in the rematch. The net difference is -0.1 percent with an estimated standard error of 0.3 percent. In Washington State, the rate of erroneous enumerations is 1.0 percent in the production and 1.2 percent in the rematch. The net difference is -0.2 percent with an estimated standard error of 0.3 percent.

Before reconciliation, the rematchers assigned more unresolved codes than the production. With reconciliation, more P-sample cases were resolved, but fewer E-sample cases were resolved. However, there were no big shifts from any one category to another.

4. CONCLUSIONS

Since we do not have complete results from Phase I and Phase II, we can make only limited conclusions at this time. However, we can make one important conclusion about matching for nonmovers. The personnel error in the matching operation for nonmovers was under control. The matching procedures were applied consistently. Evidently the careful scrutiny of the quality control operation was effective, especially since all cases were sent through quality control. The adjudication of codes assigned independently by two different matchers was probably the key to the consistency of the results.

We believe these results imply that the matching procedures can be applied consistently for nonmovers processing in 1990. We do not expect to achieve such low error rates in 1990. For 1988, we had only one office and 100 percent quality control, neither of which apply to 1990. However, these results imply that personnel error for nonmovers can be controlled. We need the results of Phase II before we can draw any conclusions about procedure error for nonmovers.

Movers were more difficult to match in 1988 and will be more difficult in 1990. The matching operation for movers is more complicated with more possibilities for error. We will be reviewing data for movers from Phase I and Phase II.

REFERENCES

Biemer, P.P. (1988) "Modeling Matching Error and Its Effects on Estimates of Census Coverage Error" Survey Methodology, 14, 117-134.

Corby, C., and Mulry, M. (1988) Memorandum to K.M. Wolter, Subject: Matching Error Pilot Study. Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

Felligi, I.P., and Sunter, A.B. (1969) "A Theory for Record Linkage" Journal of the American Statistical Associaiton, 64, 1183-1210.

Hansen, M.H., Hurrwitz, W.N., and Pritzker, L. (1964) "The estimation and interpretation of gross differences and simple response variance." In Contribution to Statistics (Ed. Cr. R. Rao) Oxford: Pergamon Press, 111-136.

Seltzer, W., and Adlakta A. (1974). On the Effect of Errors in the Application of the Chandrasekaran - Deming Techniques, Reprint 14. Laboratory for Population Statistics, University of North Carolina.

Scheuren, F., and Oh, H.L., (1985). Fiddling Around with Nonmatches and Mismatches. Proceedings of the Workshop on Exact Matching Methodologies, Arlington, Virginia.

Wolter, K.M. (1983) Affidavit, Mario Cuomo et al. vs. Malcolm Baldrige et al. U.S. District Court, Southern District of New York, 30 Civ. 4550.

* This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

Table 3.1. Results of Rematch Study: P Sample, Phase I
St. Louis Reconciled
Nonmovers

Original Results	Results of Rematching: Match	Not Match	Unresolved	Total
Match	2,667	7	8	2,682
Not Match	9	427	30	466
Unresolved	0	7	20	27
Total	2,676	441	58	3,175

Table 3.2. Results of Rematch Study: P Sample, Phase I
East Central Missouri (Unreconciled)
Nonmovers

Original Results	Results of Rematching: Match	Not Match	Unresolved	Total
Match	1,760	0	1	1,761
Not Match	2	109	4	115
Unresolved	0	0	2	2
Total	1,762	109	7	1,878

Table 3.3. Results of Rematch Study: P Sample, Phase I
Washington State (Unreconciled)
Nonmovers

Results of Rematching

Original Results	Match	Not Match	Unresolved	Total
Match	1,212	9	0	1,221
Not Match	1	66	16	83
Unresolved	0	6	12	18
Total	1,213	81	28	1,322

Table 3.4. Results of Rematch Study: E Sample, Phase I
St. Louis Reconciled
Nonmovers

Results of Rematching

Original Results	Correct Enumeration	Erroneous Enumeration	Unresolved	Total
Correct Enumeration	3,162	19	48	3,229
Erroneous Enumeration	14	49	16	79
Unresolved	8	32	171	211
Total	3,184	100	235	3,519

Table 3.5. Results of Rematch Study: E Sample, Phase
East Central Missouri Unreconciled
Nonmovers

Results of Rematching

Original Results	Correct Enumeration	Erroneous Enumeration	Unresolved	Total
Correct Enumeration	2,179	27	17	2,223
Erroneous Enumeration	15	31	2	48
Unresolved	3	1	22	26
Total	2,197	59	41	2,297

Weighted to E-sample totals.

Table 3.6. Results of Rematch Study: E Sample, Phase I
Washington State (Unreconciled)
Nonmovers

Results of Rematching

Original Results	Correct Enumeration	Erroneous Enumeration	Unresolved	Total
Correct Enumeration	1,431	4	7	1,442
Erroneous Enumeration	1	6	7	14
Unresolved	2	1	31	34
Total	1,434	11	45	1,490

Table 3.7. Results of Rematch Study: P Sample, Phase I
St. Louis Reconciled
Nonmovers

Original Results	Results of Rematching: Enumerated	Not Enumerated	Unresolved	Total
Enumerated	2,680	2	0	2,682
Not Enumerated	3	460	3	466
Unresolved	2	4	21	27
Total	2,685	466	24	3,175

Table 3.8. Results of Rematch Study: P Sample, Phase I
East Central Missouri Reconciled
Nonmovers

Original Results	Results of Rematching: Enumerated	Not Enumerated	Unresolved	Total
Enumerated	1,761	0	1	1,761
Not Enumerated	1	114	0	115
Unresolved	2	0	0	2
Total	1,764	114	0	1,878

Table 3.9. Results of Rematch Study: P Sample, Phase I
Washington State Reconciled
Nonmovers

Original Results	Results of Rematching: Enumerated	Not Enumerated	Unresolved	Total
Enumerated	1,218	3	0	1,221
Not Enumerated	2	81	0	83
Unresolved	4	10	4	17
Total	1,224	94	4	1,322

Table 3.10. Results of Rematch Study: E Sample, Phase I
St. Louis Reconciled
Nonmovers

Original Results	Results of Rematching: Correct Enumeration	Erroneous Enumeration	Unresolved	Total
Correct Enumeration	3,218	5	6	3,229
Erroneous Enumeration	3	70	6	79
Unresolved	2	2	207	211
Total	3,223	77	219	3,519

Table 3.11. Results of Rematch Study: E Sample, Phase I
East Central Missouri Reconciled
Nonmovers

Original Results	Results of Rematching: Correct Enumeration	Erroneous Enumeration	Unresolved	Total
Correct Enumeration	2,212	4	7	2,223
Erroneous Enumeration	2	45	1	48
Unresolved	1	0	25	26
Total	2,215	49	33	2,297

Table 3.12. Results of Rematch Study: E Sample, Phase I
Washington State Reconciled
Nonmovers

Original Results	Results of Rematching: Correct Enumeration	Erroneous Enumeration	Unresolved	Total
Correct Enumeration	1,437	3	2	1,442
Erroneous Enumeration	0	14	0	14
Unresolved	0	0	34	34
Total	1,437	17	36	1,490