# THE MATCHING ERROR STUDY FOR THE 1990 POST ENUMERATION SURVEY

**Mary C. Davis, Mary Mulry, and Randall Parmer, U.S. Bureau of the Census**
**Paul Biemer, Research Triangle Institute**
**Mary C. Davis, U.S. Bureau of the Census, Rm 1653-3, Washington, D.C. 20233**

KEY WORDS: Nonsampling Error, 1990 Coverage Evaluation, Capture-Recapture

## 1. INTRODUCTION

The U.S. Bureau of the Census conducted a Post Enumeration Survey (PES) to evaluate the census coverage error in the 1990 Decennial Census. This paper discusses the Matching Error Study (MES), an evaluation of the processing of the data from the PES following the census.

The PES was really two samples: a) the E sample consisting of census enumerations which measured erroneous enumerations, and b) the P sample selected independently of the census which measured census omissions (Hogan, 1991). The same blocks were selected for both the E sample and the P sample, resulting in overlapping samples. The two samples were used in dual system estimation to produce an estimate of the census coverage error.

Matching was a critical part of the 1990 PES. After the interviewing was completed in July, 1990, a match to the census was conducted to determine if the P-sample respondents were in the census. There are two basic types of errors which may occur as a result of matching -- random or systematic. Systematic errors are a particular concern since these errors may be associated with particular geographic areas and/or demographic population groups. The effect of matching error is that false nonmatches will result in an overstatement of the actual coverage error. False matches, on the other hand, will tend to understate the level of the actual coverage error.

In the 1980 Post Enumeration Program (PEP) that followed the 1980 Decennial Census, matching error was one of the most important sources of nonsampling error affecting the dual system estimate (DSE) of 1980 Census undercount (Wolter, 1983). In the 1990 PES, a number of steps were taken to reduce the error in the matching process.

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 Census Bureau's dual system estimator for an area or domain is given by:

$$\hat{N} = \frac{CN_P}{M}$$

where CEN is the size of the original enumeration for the area or domain, II is the number of imputed persons, UM is the estimate of the unmatchable census enumerations, EE is the estimate of the number of erroneous enumerations in the original enumeration, and C = CEN - II - UM - EE. $N_P$ is the estimate of the total population from the P sample universe and M is the estimate of the number of "matchable" persons in both the census and the P sample. In what follows, we refer to the combination of EE and UM (EE + UM) as EE+. The goal of the MES is to evaluate the matching error in two DSE components, EE+ and M.

Section 2 discusses the methodology for estimating the error in EE+ and M. Section 3 describes the design of the MES. The results of the data analysis are reported in Section 4. Section 5 gives a summary and conclusions.

## 2. METHODOLOGY

The major focus of the MES is the estimation of the bias in EE+ and M. That is, suppose we were able to develop a "perfect" matching system - i.e., one in which no errors were made in classifying a case as an "EE+" or "not an EE+" and as an "M" or "not an M." Let $EE+_T$ denote the value of EE+ and let $M_T$ denote the value of M from the perfect system. Let $EE+_P$ and $M_P$ denote the values of EE+ and M from the PES production system.

The bias in the production system for EE+ is $B(EE+) = E(EE+_P - EE+_T)$ and for $M_P$ is $B(M) = E(M_P - M_T)$ where $E(\cdot)$ denotes the expectation taken over all possible samples and all appropriate nonsampling error distributions. Estimates of matching error biases are only as good as the so-called perfect matching system that produces the quantities $M_T$ and $EE+_T$, i.e. the rematching system. In reality, no rematching system no matter how well designed can be expected to produce EE+ and M having no misclassification error biases. However, if the errors in $EE_T$ and $M_T$ are small relative to the size of B(EE+) and B(M), respectively, then estimates of B(EE+) and B(M) based upon the rematching system should still be useful for evaluating the production matching components.

The bias in the DSE is directly affected by the bias in EE+ and M, where $M = N_p - NM$ and NM is the estimate of the number of nonmatchable persons in the P sample. To see this, let RDR denote the "relative difference rate" defined by RDR = (production - rematch)/rematch. Thus,

$$RDR(M) = \frac{M_P - M_T}{M_T} \qquad (1)$$

and

$$RDR(EE+) = \frac{EE+_P - EE+_T}{EE+_T} \qquad (2)$$

Since the numerator of the RDR is an estimator of the bias B(M) (for (1)) and B(EE+) (for (2)), the RDR is a measure of the relative bias. It can be shown that the relative bias in $\hat{N}$, denoted by $RB(\hat{N})$, is given by

$$RB(\hat{N}) \approx E\left[\frac{RDR(C) - RDR(M)}{1 + RDR(M)}\right] \qquad (3)$$

where

$$RDR(C) = \frac{C_P - C_T}{C_T}$$

for $C_P$ = CEN-II-$EE+_P$ and $C_T$ = CEN-II-$EE+_T$.

It can be seen from expression (3) that a positive bias in $EE+_P$ (i.e., $RDR(C) < 0$) will cause a downward bias in $\hat{N}$ (i.e. $RB(\hat{N}) < 0$), ignoring the effect of RDR(M). Likewise, a positive bias in $M_P$ (i.e. $RDR(M) > 0$) also will cause a downward bias in $\hat{N}$, ignoring the effect of the bias in $EE+_P$ on RDR(C). If RDR(C) and RDR(M) have like signs, then their affects are somewhat offsetting and $RB(\hat{N}) = 0$ when RDR(C)=RDR(M). Note that in the P-sample analysis in this report, we examine RDR(NM) instead of RDR(M) so that the error rates for P-sample matching are on the same order of magnitude as the E-sample error rates. A positive RDR(NM) implies that $\hat{N}$ is biased upward, ignoring the effect of RDR(EE+).

Ultimately, every case in the E sample is classified as either a correct enumeration, CE, or an erroneous enumeration, EE. Every case in the P sample is classified as either a match, M, or a nonmatch, NM. However, following both the production matching operation and the MES rematch operation, an "unresolved" category also exists. Under certain circumstances, a case may be matchable yet the information is insufficient for an accurate match. For example, an incomplete mover address is given to determine census day residence. Since the exact April 1 location is not known, the case cannot be resolved. Such cases are classified as "unresolved" or UR for both the E and P sample. Subsequently, in the computer imputation stage, the URs will be resolved. A match probability (for the P sample) and a probability of correct enumeration (for the E sample) are imputed for these cases in the PES imputation process. However, the analysis in this report is focused on the data which emerged from the production matching operation before the imputation process for the unresolved cases.

## 3. STUDY DESIGN

### 3.1 Production Matching

After the interviewing phase of the PES, the matching operation determined whether the P-sample respondents were enumerated in the census.

The matching operation occurred in two major phases: before followup (BFU) matching and after followup (AFU) matching. 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, E-sample nonmatches, P-sample unmatchables, and E-sample unmatchables.

A clerical matching operation was performed following computer matching. A computerized quality control system kept track of the codes assigned during the various steps of clerical matching. The first level of matchers, called the Clerical Matching Group (CMGs), were given rules for designating matches. The second level, called the Special Matching Group (SMGs), were given more flexibility in using their judgement to

designate matches. For quality control, a second SMG clerk "independently" matched each cluster. "Independently" means that the matcher did not have access to match codes assigned by either the CMG clerk or the first SMG clerk. The match results from the CMG clerk and the first SMG clerk were compared with the results from the second SMG clerk. Any differences in the two sets of codes were adjudicated by a PES technician who assigned a reconciled code. Clusters with a high nonmatch rate were reviewed by a Matching Review Specialist (MRS), the highest level of matcher. The MRS were given extensive training in assigning match status in the most difficult of cases.

Cases identified as requiring further information (followup) were sent to the field for resolution. After the followup, the SMG clerks used the new information and attempted to resolve the case. Two SMG clerks independently reviewed the case and when the SMG clerks disagreed, a technician adjudicated the cases and substituted a final match code. The MRS reviewed up to 15 percent of each processing office's clusters with high unresolved rates or with E-sample geocoding problems.

### 3.2 Matching Error Study Design

The MES matching was conducted in each of the seven processing offices (Albany,NY; Austin,TX, Baltimore,MD; Jacksonville,FL; Jeffersonville,IN; Kansas City,MO; and San Diego,CA) following the termination of all PES operations. A dependent rematch of the 919 block cluster sample was performed by a group of matching personnel consisting of MRS and PES technicians. (See section 3.3 for a description of the sampling procedure.) "Dependent" means that the matchers had access to the match codes assigned at every stage of the production matching. However, procedures were implemented to insure that the assignment of MES match codes was not influenced by the production matching operation by not allowing matchers to work in any processing office in which they had worked in production.

The MES matching operation was performed in two stages. During the first match, the MES matchers used all relevant materials from production to assign an MES match code to every person in the cluster. When additional information was required, cases were sent out for field followup provided the case had not gone out either in PES production or the Evaluation Followup (EFU) Study (West,1991). These procedures were implemented to insure that the results of the MES analysis would not include "data collection error" which is evaluated separately by the EFU study. As a result of the followup guidelines, only 152 cases were sent to the field, 144 were E-sample cases while only 8 were P-sample cases.

The final MES match was conducted following the followup field work using only the MRS staff. The results of the followup were used to resolve the case and assign a final MES match code. During both matching stages, clusters were 100 percent reviewed by MRS personnel to insure that all match codes were reviewed and assigned by the most highly trained matching personnel.

Following the MES matching operation in the processing offices, a supplemental review of the most difficult E-sample cases was performed by the MRS staff. "Difficult"cases were defined as census fictitious, noninterviews from followup, or incomplete information on the followup form. If these codes were assigned as the production final match code or in either the MES first match or the final match, the case was reviewed. As a result of the review, the final MES match code was changed for 310 cases. Note that only MES match codes were changed during the review.

### 3.3 Sample Design

The PES evaluation sample is a stratified systematic sample of the PES sample block clusters. The PES sampling strata were first grouped into thirteen evaluation sampling groups which were approximately the same as the thirteen evaluation poststrata. Within each sampling group, block clusters were further grouped and sorted using criteria aimed at minimizing the variances of the estimated error rates. An unequal probability sample of 919 block clusters were drawn from the 13 sampling groups. For the allocation of sample clusters to evaluation groups, an optimal allocation strategy was followed.

## 4. DATA ANALYSIS

### 4.1 Objectives

The primary objective of the MES was to evaluate the quality of the E-sample and P-sample production matching operation. Determinations as to whether or not matching error was controlled in the PES are based primarily upon results at the evaluation poststratum (EPS) levels shown in Figure 4.1. However, matching error is also examined at the processing office (PO) level and for some demographic subgroups. Recall that these estimates do not reflect the final depositions of the unresolved

cases for either the production matching or MES rematching operation.

As a rough guide, an RDR(EE+) of -0.1 or -10% indicates a positive bias in the DSE of about 0.5%, ignoring the P-sample matching error in NM. In other words, a -10% RDR(EE+) indicates that the population in the EPS would be overestimated by roughly one-half percent due to E-sample matching error. This is a very rough guide since the UR cases could change the EE+ rate when these cases are imputed. Similarly, a -10% for RDR for NM indicates that the population in the EPS would be overestimated by roughly one-half percent due to P-sample matching error.

**Figure 4.1** The 13 Evaluation Poststrata

1 Northeast, Central City, Minority
2 Northeast, Central City, Nonminority
3 U.S., Noncentral City, Minority
4 Northeast, Noncentral City, Nonminority
5 South, Central City, Minority
6 South, Central City, Nonminority
7 South, Noncentral City, Nonminority
8 Midwest, Central City, Minority
9 Midwest, Central City, Nonminority
10 Midwest, Noncentral City, Nonminority
11 West, Central City, Minority
12 West, Central City, Nonminority
13 West, Noncentral City, Nonminority + Indian

4.2 E-Sample Analysis

The estimates for the RDRs for EE+ defined by Equation (2) in Section 2 were computed for each processing office, the U.S., and the 13 EPS. A negative RDR for EE+ implies that the DSE of the total population size is biased upward.

As Table 4.1 shows, three POs have significant RDR(EE+): Baltimore, Jacksonville, and Kansas City. Each office found significantly fewer EE+ in production matching than in the MES rematching at the 95% level of confidence.

As Table 4.2 shows, only EPS 6 has a significant RDR(EE+) of -0.101. This finding is consistent with the PO findings since 78% of EPS 6's total weight is contributed by Baltimore, Jacksonville, and Kansas City. With a -10% level of bias in the number of EE+'s, the DSE would overestimate the total population in EPS 6 by approximately 0.5%, assuming there were no other compensating or contributing errors. For other EPS the levels of bias in the estimates of EE+ are too small to be detected with the MES sample. For the U.S., the RDR(EE+) is not significant.

**Table 4.1** 95% Confidence Intervals for RDR(EE+) by Processing Office

| PO | RDR | LCL | UCL |
|---|---|---|---|
| ABPO | -0.058 | -0.147 | 0.030 |
| AUPO | 0.090 | -0.042 | 0.221 |
| BAPO | -0.122 | -0.233 | -0.010 |
| JXPO | -0.118 | -0.196 | -0.039 |
| JFPO | 0.027 | -0.032 | 0.086 |
| KCPO | -0.101 | -0.176 | -0.027 |
| SDPO | 0.006 | -0.081 | 0.068 |

**Table 4.2** 95% Confidence Intervals for RDR(EE+) by Evaluation Poststrata

| EPS | RDR | LCL | UCL |
|---|---|---|---|
| 1 | -0.048 | -0.139 | 0.042 |
| 2 | -0.024 | -0.085 | 0.037 |
| 3 | 0.068 | -0.067 | 0.203 |
| 4 | -0.109 | -0.241 | 0.022 |
| 5 | -0.117 | -0.284 | 0.049 |
| 6 | -0.101 | -0.199 | -0.003 |
| 7 | -0.058 | -0.181 | 0.066 |
| 8 | 0.017 | -0.089 | 0.123 |
| 9 | -0.032 | -0.124 | 0.060 |
| 10 | -0.001 | -0.071 | 0.070 |
| 11 | 0.029 | -0.018 | 0.076 |
| 12 | 0.044 | -0.181 | 0.270 |
| 13 | -0.061 | -0.183 | 0.060 |

Matching error was also examined for seven race/hispanic origin categories shown in Figure 4.2. However, none of the RDR(EE+)s were significant. We conclude that the levels of bias in the estimates of EE+ for these categories are too small to be detected with the MES sample.

When the poststrata are regrouped by minority (EPS 1,3,5,8,and 11) and nonminority (EPS 2,4,6,7, 9,10,12,and 13), the production estimate of EE+ for nonminority was significantly smaller than the rematch estimate (RDR(EE+) = -0.052),indicating that EE+ are underestimated by about 5% for this subgroup. The effect of this bias on the DSE of $\hat{N}$ is small, resulting in an upward bias in $\hat{N}$ of about one-fourth of one percent.

**Figure 4.2** Race/Hispanic Origin Categories

1 White, Nonhispanic, Others
2 Black, Nonhispanic
3 Hispanic
4 American Indian, Nonhispanic
5 American Indian, Hispanic
6 Asian, Nonhispanic
7 Asian, Hispanic

4.3 P-Sample Analysis

The estimates for the RDRs for NM were computed for each processing office, the U.S., and the 13 EPS. A positive RDR for NM implies that the DSE of the total population size is biased upward.

As shown in Table 4.3, Albany is the only PO to have a significant level of matching error bias for NM. With a positive RDR(NM) of 0.085, we see that production matching overestimated the number of nonmatches.

**Table 4.3** 95% Confidence Intervals for RDR(NM) by Processing Office

| PO | RDR | LCL | UCL |
|---|---|---|---|
| ABPO | 0.085 | 0.016 | 0.154 |
| AUPO | -0.008 | -0.047 | 0.031 |
| BAPO | 0.049 | -0.120 | 0.218 |
| JXPO | 0.007 | -0.036 | 0.050 |
| JFPO | 0.036 | -0.040 | 0.112 |
| KCPO | 0.012 | -0.039 | 0.063 |
| SDPO | 0.002 | -0.080 | 0.084 |

In Table 4.4, we see that EPS 1 and EPS 8 have significant RDR(NM). The point estimate of the bias in NM for EPS 1 is 0.067 indicating that the number of NMs for EPS 1 were overestimated by 6.7%. The potential impact of this bias on the DSE is to overestimate the population in EPS 1 by approximately 1.3%, assuming no other compensating or contributing errors. In EPS 8 the point estimate of the bias in NM is 0.044. The potential impact of this bias on the DSE, assuming no other errors, is to overestimate the population in EPS 8 by 0.7%.

Approximately 76% of the population in EPS 1 is contributed by Albany. Thus, we can surmise that the significant positive bias in NM observed in EPS 1 is due to the Albany PO. For EPS 8, the population is split between the Jeffersonville PO (with 85%) and the Kansas City PO (with 15%), neither of which have significant RDR(NM)s.

The only significant RDR for race/Hispanic origin domains is for the Black population with RDR(NM) = 4.5%. The potential impact of this estimated relative bias on the DSE is to overestimate the total population of Blacks by approximately 0.7%.

**Table 4.4** 95% Confidence Intervals for RDR(NM) by Evaluation Poststrata

| EPS | RDR | LCL | UCL |
|---|---|---|---|
| 1 | 0.066 | 0.013 | 0.119 |
| 2 | 0.049 | -0.063 | 0.161 |
| 3 | 0.004 | -0.051 | 0.059 |
| 4 | 0.151 | -0.012 | 0.314 |
| 5 | 0.028 | -0.031 | 0.087 |
| 6 | -0.020 | -0.073 | 0.033 |
| 7 | -0.018 | -0.065 | 0.029 |
| 8 | 0.044 | 0.003 | 0.085 |
| 9 | -0.012 | -0.069 | 0.045 |
| 10 | 0.055 | -0.055 | 0.165 |
| 11 | 0.040 | -0.048 | 0.128 |
| 12 | -0.023 | -0.150 | 0.104 |
| 13 | -0.019 | -0.088 | 0.050 |

When the poststrata are regrouped by minority and nonminority, neither group was found to have a significant RDR at the 5% level of significance.

A subgroup of the P sample are movers; either persons who moved into the sample block after census day or persons who had an alternate address on census day. The address given for census day is searched to determine whether the person is counted in the census. The mover analysis shows that none of the RDR(NM)s for the seven POs is significantly different from 0 at the 5% level. However, except for Jacksonville, the general trend for movers and the overall P sample is the same - overestimation of NM. There was no evidence of a matching quality differential for movers. However, the standard errors on the RDR(NM)s for movers are quite large indicating low precision in the estimates.

## 5. SUMMARY AND CONCLUSIONS

### 5.1 E Sample

Several significant results were found for the E sample when the MES data are weighted to the total population. Baltimore, Jacksonville and Kansas City found significantly fewer EE+'s in the production operation than the MES rematch. When the MES data are regrouped by evaluation poststrata, only EPS 6--the South, Central City, Minority poststratum, was found to have significantly fewer EE+'s in production than in the MES. There is also evidence that EE+'s are underestimated in production for nonminorities. However, estimates of RDR(EE+) by Race/Hispanic Origin are not significant. Note that for estimating the bias in the number of EE+'s, the evaluation sample produced large standard errors. Confidence intervals (95% level) for RDR(EE+) averaged 12 percentage points in length. Design effects ranged from 1 to 7 with an average of 4.

### 5.2 P Sample

Several conclusions are made about P sample matching error. Almost all POs had difficulty with matching the Central City, Minority persons in the P sample. The production matching operation produced significantly more nonmatches for these areas than did the rematch. Overall, however, the production matching operation did not differ significantly from the rematch operation in the number of P-sample nonmatches. Only Albany with an RDR(NM)=6.4% was significant. However, across all POs the general trend was to overestimate nonmatches. The population sizes in EPS 1 and EPS 8 would be significantly overestimated by the DSE if P-sample matching error were the only error affecting the DSE. The magnitudes of the biases in the population sizes due to matching error are approximately 1.3% for EPS 1 and 0.7% for EPS 8. Nonmatches for Blacks were overestimated by about 4.5%. This equates to a potential positive bias in the DSE of the total Black population of approximately 0.7%. Note, for estimating the bias in the number of NMs, the evaluation sample produced very large variances. Confidence intervals (95% level) averaged 16 percentage points and for total population weighted estimates, some design effects were as large as 13 (Baltimore). The average design effect was 4.

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

Note: Due to space limitations, analysis of the impact of matching error on the P and E sample population estimates was removed. A complete paper is available from the contact author.


## ACKNOWLEDGEMENTS

The authors wish to thank Courtney Ford and Lynn Harnois for computer support and Chris Moriarity for variance estimation. Special thanks to Martha Sutt and Dan Childers for helping to organize and guide the matching operations in the processing offices.


## REFERENCES


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

Hogan, H.R. (1991) "1990 Post-Enumeration Survey Operations and Results" Proceedings of the Section on Social Statistics, American Statistical Association.

West, K., Mulry, M., Parmer, R., and Petrik, J. (1991) "Address Reporting Error in the 1990 Post-Enumeration Survey" Proceedings of Survey Research Section, American Statistical Association.

Wolter, K.M. (1983) Affidavit. Mario Cuomo *et al. vs* Malcolm Baldridge *et al.* U.S. District Court, Southern District of New York, 80 Civ. 4550.