Statistical Science
1994, Vol. 9, No. 4, 458–537

The 1991 Census Adjustment: Undercount or Bad Data?

Leo Breiman

Abstract. The question of whether to adjust the 1990 census using a capture–recapture model has been hotly argued in statistical journals and courtrooms. Most of the arguments to date concern methodological issues rather than data quality. Following the Post Enumeration Survey, which was designed to provide the basic data for adjustment, the Census Bureau carried out various evaluation studies to try to determine the accuracy of the adjusted counts as compared to the census counts. This resulted in the P-project reports, which totaled over a thousand pages of evaluation descriptions and tables. Careful scrutiny of these studies together with auxiliary sources of information provided by the Census Bureau is used to examine the issue of whether the data gathered in the Post Enumeration Survey can provide reliable undercount estimates.

Key words and phrases: Census, Post Enumeration Survey, nonsampling error, undercount.

1. INTRODUCTION AND SUMMARY

To give the setting for this paper, we begin with a simple example. Consider a project undertaken to find the total fish population of a large pond. Efforts are made to catch all of the fish and paint a red X on their backs. In total 10,000 fish are caught and marked. To see if this effort gave a complete count, 100 fish were later caught and examined. Of these, 98 had X's on their backs, and two did not. If the recapture (second catch) is done at random with each fish in the pond having the same recapture probability, and if the population of the pond stays the same between the initial catch and the recapture, then an approximately unbiased estimate for the total pond population is 10,204 and there is an estimated undercount of 2.0%. Such estimates are called capture–recapture estimates.

However, suppose that subsequent study revealed that there may have been X's on the backs of the two fishes, but perhaps the X's had not been well painted on to begin with or that the examination had not been well carried out. Instead of there being a 2% undercount, what may be true is that there was 2% bad data, or perhaps there was a 1% undercount and 1% bad data. The question of how much of the data is bad is fundamental to knowing how accurate the undercount estimate is.

The effort to adjust the 1990 census for undercount was arduous. It consisted of following the census with the Post Enumeration Survey (PES), covering 380,000 persons; matching the census and PES records; and then computing capture–recapture estimates of the undercount. We examine two related questions. First, what are the sources of errors in the PES and matching, and how big are these errors? Second, what is the effect of the errors on the undercount estimates? In the end, our objective is to see whether, in retrospect, the census adjustment proposed in 1991 is statistically justifiable.

Along the way, we will try to give full references to the relevant published literature. Many of the important documents are unpublished internal Bureau of the Census reports or reports generated by other government agencies. We will cite these, noting that they are available on request from the proper agency. Some references are included that provide background information but do not have an immediate connection to the issues raised in the text. These are described in Section 7.

1.1 Background

The issue of whether to adjust the 1990 census using a capture–recapture model has been one of the most highly publicized and important statistical issues of the past decade. It has serious political and economic consequences. Census counts are used to apportion congressional and legislative seats and to

Leo Breiman is Professor, Department of Statistics, University of California, Berkeley, California 94720.

distribute tens of billions of dollars that flow from the national government to states, counties and cities.

Part of the long-term Census Bureau planning for 1990 was to follow the census with a Post Enumeration Survey (PES) covering over 150,000 households, to match census persons to PES persons and, using capture–recapture assumptions, to compute adjusted estimates of the census counts at the national, state and local levels.

The Department of Commerce decided in 1987 not to issue official 1990 census counts that were statistically adjusted. The estimates based on the PES data and matching would be used only to "provide a careful evaluation of the coverage of the 1990 Census" (Hogan and Wolter, 1988). However, under legal pressure from a group of cities and states who favored adjustment, the Department of Commerce in July 1989 agreed to initiate a new decision making process (see Department of Commerce, 1991b, Section 4, Appendix 1).

In March 1990, prior to the census and the PES, the Department of Commerce published guidelines to follow in deciding whether the adjusted census counts would be officially adopted (Department of Commerce, 1989). The most important, from a statistical point of view, was Guideline 1: "The Census shall be considered the most accurate count of the population of the United States, at the national, state, and local level, unless an adjusted count is shown to be more accurate."

The census interviews of households not returning their questionnaires by mail took place in May–July 1990. The PES was carried out in July–August 1990. The adjusted counts for the nation, states and larger counties and cities were released in June 1991 and showed an estimated national undercount of about 2%—5 million people (Department of Commerce, 1991a). The largest estimated undercounts were, as expected, among minorities in central cities.

On July 15, 1991, the Secretary of Commerce made the decision not to adjust (Department of Commerce, 1991b) and a group of cities and states (including, e.g., California, Los Angeles, Atlanta, New York, Florida and Texas) sued to force adoption of the adjusted numbers. Statisticians testifying for the plaintiffs were Barbara Bailar, Eugene Ericksen, Stephen Fienberg, John Rolph, John Tukey and Kirk Wolter. Testifying for the government were Peter Bounpane, Robert Fay, David Freedman, Paul Meier and Kenneth Wachter. Leo Breiman assisted the defendants. The testimony ended on May 28, 1992.

On April 13, 1993, the court issued its holding that "the decision against adjustment shall not be disturbed,..." (U.S. District Court, 1991). The holding was based on the grounds that, since reasonable statisticians could differ on the merits of adjustment, "the Secretary's decision not to adjust the 1990 Census count was neither arbitrary nor capricious." The court did not base its ruling on the relative accuracy of the adjusted versus nonadjusted counts, and controversy over this latter issue will probably continue in the pages of statistical journals for years to come (with reasonable statisticians on both sides).

1.2 The Adjustment Method

The adjustment undertaking was extraordinarily complex. Over the decade preceding the 1990 census, much research had been concentrated on this issue by statisticians inside and outside of the Census Bureau. Four rehearsals aimed at uncovering deficiencies in the methodology were carried out. The first was in Mississippi in 1986, the second in Los Angeles in 1986, the third in North Dakota in 1987 and the last in Missouri and Washington in 1988. Numerous published papers, committee reports and thousands of pages of internal Bureau of the Census studies dealt with various issues involved. The reference section lists many of these.

The method used in 1990 to estimate adjusted counts at all levels down to the census block consisted of defining 1,392 poststrata on the basis of the following: age; sex; race or ethnicity; renter or owner; place type (i.e., central city, nonincorporated area, etc.); and geographic location. For instance, one poststratum is male, ages 10–19, black renter, central city, New England. All persons in the population, with some insignificant exceptions (Undercount Steering Committee, 1991), are in one of the poststrata.

Then the capture–recapture idea is applied, that is, there is an original survey (the census), followed by another survey (the PES). A matching is done to see how many of the persons found in the second survey were also found in the first. The census count is corrected for erroneous enumerations. In each poststratum, the corrected census count, the PES count and the number of matches are used to compute a capture–recapture estimate of the poststratum population. The population estimate divided by the original census count is the raw adjustment factor for the poststratum.

The population estimate is called the Dual System Estimate (DSE) (Wolter, 1986b). We stretch terminology by referring to the adjustment factors and the undercounts also as DSE estimates. Now we can give an explanation for the poststrata definition. The a priori belief was that the stratification used created population pools (each poststratum) having approximately equal recapture probabilities, thus validating the capture–recapture assumptions.

Statistical techniques were then used to produce a set of smoothed adjustment factors having smaller variances than the raw factors. The estimated variance–covariance matrix of the raw adjustment factors was "presmoothed," that is, regressed against some explanatory variables including, for example, indicators for gender, age and minority. Then the raw adjustment factors were smoothed using a Bayesian version of linear regression and the presmoothed variance–covariance matrix, with the variables in the regression selected using a "best subsets" method and Mallows' C_p. See Freedman et al. (1993) for a more detailed description.

Adjusted counts for each local unit are computed by slicing it into poststrata, adjusting each poststratum for its undercount and then recombining. For instance, suppose that a certain city has a census count of 10,000 in a poststratum and that this poststratum has an estimated adjustment factor of 1.035. Then the adjusted city count in this poststratum is $10{,}000 \times 1.035 = 10{,}350$. The overall adjusted count for the city is obtained by adjusting the counts in all poststrata intersecting the city population and adding these up.

The extent of the error introduced into the DSE estimates by using common adjustment factors over potentially diverse populations is a subject of controversy (Freedman and Wachter, 1994). For instance, the adjustment factor for the poststratum male, ages 10–19, black renter, central city, New England is used to adjust the counts for all central cities in New England. There is also controversy about the statistical validity of the smoothed adjustment factors and their estimated variances and covariances (Freedman et al., 1993). This paper, however, deals only with the issue of the errors in the DSE estimates attributable to errors in the PES data and the matching process.

1.3 Summary

This paper is laid out in two parts. First, Sections 2 and 3 describe the steps leading to the DSE estimates and the evaluation procedures and documents. Second, Section 4 discusses sensitivity of the undercount estimates to errors, and Sections 5 and 6 give an analysis of the errors and potential errors. Here is a brief summary and conclusion as an introduction and a road map.

The PES covered 380,000 people in 169,000 households. These households were in 5,290 block clusters selected using a stratified random design. Then efforts were made to match the PES data to the census data from the same blocks. This was a multistep process involving the reinterview of some households where better information was needed. Cases that could not be resolved into a match category were sent to an imputation process. (See Section 2.)

After the PES was completed, evaluation data was gathered through a rematching study, some field reinterviews and examination of quality assurance records. The evaluation results were summarized in documents called P-project reports. These studies form the main source of information for the error discussions in this paper. Due to the smaller sample sizes in the evaluation data, results were presented only at highly aggregated levels, that is, either as weighted to national figures or to the nation divided into 13 large evaluation strata defined by minority versus nonminority, by central city versus non–central city and by geographic region, Northeast, South, Midwest, West (see Section 3.5, Table 4). (See Section 3.)

To understand the subsequent error analysis, it is first necessary to understand that small errors in the PES can result in large errors in the undercount estimates. The DSE undercount estimates are computed from estimates of the census counts corrected for erroneous enumerations, the PES counts and the number of matches. Then, for instance, errors of 0.5% in the match count and in the PES count can together lead to a 50% error in the undercount estimate. Also, because of the uneven weighting due to the sampling design, mistakes in a handful of people may have large effects on undercount estimates. (See Section 4.)

There are many potential sources of errors. For instance, the number of matches can be erroneous because of mistakes made by the matching teams or because some of the data is missing, unreliable or fabricated. Persons moving in or out of the sample blocks after census day (April 1, 1990) can be an error source. Deciding just which housing units are in the designated blocks is not always simple, and mistakes can result in errors in the DSE estimates. The final imputation process is another potential error source. The P-studies are examined to see what evidence there is concerning the magnitudes of the errors and data quality. (See Section 5.)

Some error estimates can be obtained from the evaluation data aggregated to the national level and to the level of the 13 evaluation strata. The DSE national undercount estimate was 2.1%. The Census Bureau Total Error Report [P16] estimated that various errors led to an upward bias of 0.7% in the DSE estimate. Correcting the initial estimate for these errors gives an estimate of 1.4%.

After the Secretary's decision a coding error in processing the PES data was discovered that lowered the undercount estimate by 1,000,000 persons. The rematching of some suspect blocks led to a 250,000-person decrease. As a result, the Census Bureau lowered the DSE estimate to 1.6% with an estimated upward bias of 0.7% (Mulry, 1992b). Correcting for this bias drops the estimate to 0.9%. Our scrutiny of the P-studies showed some errors not included in the total error report that lower the national estimate to 0.4%. Other problems could lower the figure even more. (See Section 6.)

Thus, using the Census Bureau estimates, 55% of the DSE national undercount estimate is due to bad data or processing errors; our estimate is at least 80%. Also, there is an uneven distribution of errors. The corrections downward in the five minority evaluation strata are considerably larger than in the nonminority strata (see Section 6.3, Table 16). Some individual evaluation strata were affected more than others. For instance, the minority, central-city stratum in the South has a DSE undercount estimate of 5.7%. The corrected estimate is 1.5%.

Our error analysis does not always agree with the Census Bureau analysis, and the differences are pointed out in Section 6 and the Appendix. The most recent account of the Census Bureau analysis is Mulry and Spencer (1993). This is one of a collection of articles about the 1990 undercount published in the September 1993 issue of the *Journal of the American Statistical Association* (*JASA*). Note, however, that this article does not correct for the coding error or other errors found by the Census Bureau in late 1991 and early 1992 (see Hogan, 1993, in the same collection). Therefore the error estimates, the loss function analysis and conclusions in the article do not reflect current knowledge.

1.4 Conclusions

The PES data are not reliable enough to give accurate undercount estimates. To a substantial extent the 1991 DSE undercount estimates are artifacts reflecting data quality. The largest part of the original undercount estimate is due to bad data and processing error—80% on the national level. It is difficult to deduce how much more of the undercount estimate is similarly affected.

Because of the relatively small sample sizes in the evaluation data, results are broken out by the 13 evaluation strata rather than the 1,392 poststrata. Thus, it is not known what effect the error rates and decreases in the undercount estimates indicated by the evaluation data would have on the undercount estimates at the county, city, and state level. However, the fact that 80% of the undercount estimate aggregated to the national level is due to poor data quality indicates that the DSE estimate state and local adjustments are largely reflections of bad data in an unpredictable pattern.

TABLE 1
PES household interview outcomes

Interview with household member	93.7%
Proxy interview	4.2%
Noninterview	1.6%
Out-of-scope	0.5%

The results of this study should not be taken to mean that I believe that the true 1990 undercount is as low as 0.4% or even 0.9%. My focus was on whether the 1990–1991 DSE estimate process produced reasonable estimates of the true undercount. To that, my answer is no; there were simply too many sources of error. The accuracy necessary in dozens of diverse areas to keep the total error within the requisite bounds was simply not attainable. It is not attainable now and probably not in the future.

2. PRODUCTION OF THE UNDERCOUNT ESTIMATES

The procedure that produced the raw DSE undercount estimates for the 1,392 poststrata can be put into four phases. The first, which we call the initial PES, is similar to many other sample surveys. Addresses are listed and interviewers go from one household to the next. The PES was carried out in July–August 1990.

The next three phases are unique: they consist of the matching, the follow-up interviews to obtain additional information on hard-to-match cases (November–December 1990) and more matching followed by imputation. Planning for the PES and the subsequent phases is discussed in Anolik (1990), Biemer and Stokes (1989), Childers et al. (1987), Hogan (1989), Hogan and Wolter (1988) and Wolter (1987a). Descriptions of operations and results are given in Department of Commerce (1991b, Section 4), in Hogan (1993), and in Undercount Steering Committee (1991).

2.1 The Initial PES

The Post Enumeration Survey covered 380,000 people in 169,000 households. These households were in 5,290 block clusters selected using a stratified random design. Woltman, Alberti and Moriarty (1988) summarize the sample design of the PES, and Hogan (1990) gives an overall description.

The first phase of the PES consisted of listing addresses in the designated blocks. Then interviewers covered the address lists. The questionnaire covered name, age, sex, race or ethnicity, owner or renter, marital status, relation to head of household and address as of April 1, 1990 (census day). A summary of the outcomes of household interviews is given in Table 1 (Diffendal and Belin, 1991, Table 3.1).

"Proxy interviews" are interviews with nonhousehold members such as neighbors, apartment managers, landlords or other knowledgeable respondents. "Out-of-scopes" are persons that do not belong in the PES sample for a variety of reasons (i.e., home on a military leave, visiting for the weekend, duplicate record, etc.).

2.2 Matching

The persons found by the PES in the designated block clusters are referred to as the P-sample. The persons found by the census in the same block clusters constitute the E-sample. The next phase of the PES undercount estimation procedure consisted of attempting to match E-sample census records to P-sample PES records. The Census Bureau has done research on improving matching techniques for over a decade, and the methods used on the 1990 data were comprehensive and well rehearsed.

The matching was done in three phases. The first phase consisted of computer matching (Jaro, 1989). This matched 75% of the records. In the next phase, the records unmatched by the computer went through each of two independent tracks. One track consisted of two matching teams; the other, of one. In the third phase, an adjudication team assigned a match status code to all records on which the two previous tracks had disagreed. There was also a special team assigned to double-check match status codes in suspect data.

The match codes are complex and differ for the P- and E-samples. They fall into three basic categories. For the P-sample, these are match, nonmatch and unresolved. For the E-sample, these are correctly enumerated, erroneously enumerated and unresolved. Persons could be erroneously enumerated in the census for a variety of reasons; for example, if they were born after April 1, 1990, or died before this date, or if they were enumerated at the wrong address or were enumerated more than once.

We have not been able to find a single document giving a detailed but readable description of the matching rules, codes and procedures. What exists are lists of match codes and lengthy, detailed manuals of instructions for the matching teams. The best brief summaries are in Department of Commerce (1991b, Section 4) and in West, Corby and Van Nest (1989). A more extensive description of the similar matching procedures used in the 1988 rehearsal is in Childers and Hogan (1989a).

TABLE 2
PES follow-up household interview outcomes

	P-sample	E-sample
Interview with household member	81.6%	79.2%
Proxy interview	17.0%	19.6%
Noninterview or out-of-scope	1.4%	1.3%

2.3 The Production Follow-up

After the initial matching procedure, many persons were unresolved or unmatched. To get additional information about these cases, a follow-up survey was done. This included the following:

1. all people in the E-sample not matched to the P-sample;
2. all P-sample whole household nonmatches;
3. proxy interview P-sample partial household nonmatches.

A total of 47,000 households were sent to follow up [P2, Table 3.2]. The outcomes of household interviews [P2, Tables 6 and 7] are given in Table 2.

2.4 Imputing the Unresolved

After the new information gathered in the production follow-up was used in the matching procedure, 12,500 persons in the E- and P-samples remained unresolved [P1]. Statistical models were used to impute how many of the unresolved P-sample persons should be assigned as matches in each poststratum and how many of the unresolved E-sample persons should be assigned as correctly enumerated in each poststratum. These models are described in Belin et al. (1993) and Diffendal and Belin (1991).

Following the imputation procedures, the numbers in each poststratum needed to estimate the poststratum adjustment factor are on hand: the census count; the P-sample count; the number correctly enumerated by the census; and the number of matches. The output consists of the raw adjustment numbers for each poststratum (see Section 4.1).

3. THE EVALUATION DATA AND STUDIES

The Census Bureau planned and carried out an evaluation of the process leading to the DSE estimates. Sources of potential error were categorized, and projects were designed to gather and/or analyze data in order to estimate the magnitudes of the errors. The evaluation data gathering and analysis were mainly done following the completion of the

PES. The results are detailed in 22 evaluation project reports, referred to as the P-project reports.

Data for the evaluation projects (except [P13] and [P18]) come from three sources. The first consists of records from the quality control and quality assurance procedures that were ongoing during the PES and matching. The second is a rematching study carried on using a subsample of the P- and E-records but different matching teams. The third, and most important, is an evaluation follow-up (EFU) which reinterviewed a subsample of the P-sample and E-sample persons.

The narratives and tables of the P-project reports form the basis for most of the error analysis in this report. For this reason we give a more detailed overview of the evaluation and the P-project reports. This consists first of a listing of the reports and other associated references. Second, a description of the evaluation data sources is given. The results in the P-project reports are usually given by aggregation into 13 evaluation strata (the definitions of these strata are given in Section 3.4).

3.1 The Evaluation Studies

All project reports were obtained from the census. Various types of error are connected with evaluation project reports as follows:

- matching errors, P-sample—[P7], [P8];
- matching errors, E-sample—[P9a], [P10];
- interviewer fabrications—[P5], [P5a], [P6];
- census day address errors—[P4];
- missing data and imputation error—[P1], [P2], [P3];
- incorrect address coding—[P11];
- correlation bias—[P13];
- errors due to late census data—[P18];
- total error summary—[P16].

The full references for the P-project reports are listed at the end of the reference section. Recently, extracts from some of these reports have appeared in proceedings volumes or journals: [P2] in Gbur (1991a); [P4] in West, Mulry, Palmer and Petrik (1991); [P5] in Tremblay, Stokes and Greenberg (1991); [P7] and [P10] in Davis, Mulry, Palmer and Biemer (1991); and [P16] in Mulry and Spencer (1991, 1993). The published articles omit much of the detail given in the parent P-project reports. An overall view of the purposes and methods of evaluation is given in Hogan (1989). Section 4 of Department of Commerce (1991b) contains executive summaries of all of the evaluation project reports, but there is no single document that gives full descriptions of all of the evaluation projects.

The Census Bureau's intention was to make the P-project reports the definitive collection of information regarding the evaluation of the PES and the subsequent undercount estimates. Almost all of the material in this report is drawn from these evaluation studies. Supplementary data about the P-sample reinterviews in the EFU and the effects of the coding error discovered in late 1991 were supplied by the Census Bureau.

3.2 Quality Control and Assurance Data

During the interviewing phase of the PES, an ongoing quality control operation was in place to confirm that the PES interviewers visited the correct households and conducted the interviews according to survey procedures, and to conduct reinterviews of questionable work. Overall, about 35% of the P-sample was reinterviewed by phone or personal visits in this quality control operation [P5]. This data was used in [P5] to estimate the extent of undetected fabricated interviews.

In production matching, at least three and up to five different teams were involved in the final decision on each record not matched by computer. Logs were kept on the intermediate decisions. These were used in [P8] to quantify the ranges of disagreement between the various teams. This data source is referred to as the quality assurance results.

3.3 The Matching Error Study

In the Matching Error Study, at each processing center a sample of P- and E-records were selected and rematched. According to the bureau, the rematching was done more carefully than the production matching. Report [P10] notes that in the rematching "all match codes... were reviewed by MRS, the most highly trained matching personnel." In production matching the matching review specialists (MRS's) reviewed only a fraction of the codes assigned. Rematching was undertaken for 71,000 P-sample cases.

The rematching was not independent of the original matching. The rematch teams had available to them all of the match codes assigned in the production matching. The Matching Error Study estimates what happens when the matching is repeated using the same matching rules, the same computer algorithm and with the previous matching information available to the rematch team. Brief descriptions are given in reports [P7] and [P10].

3.4 Evaluation Follow-up (EFU)

In this phase, carried out in February 1991, a subsample of households were reinterviewed to get additional information about Census Day address errors,

TABLE 3
EFU household interview outcomes

	P-sample	E-sample
Interview with household member	87.7%	85.2%
Proxy interview	10.4%	13.4%
Noninterview	1.9%	1.4%

noninterviews, fabricated data, imputations and so on. The interviewers had available the records of the past interviews. About 11,000 households were reinterviewed, and data collected for 27,000 persons ([P4], [P9a]). The design of the EFU is described in Reports [P3], [P4], [P5a] and [P9a]. The interview results are shown in Table 3 [P2, Tables 8, 9 and 11].

The interviewers used were more experienced than those used in the PES. To quote report [P3], "A staff composed of only current survey interviewers was used for the EFU interviewing. The interviewers hired and trained for the PES and the Census were primarily temporary employees... ." Report [P4] states that the team who matched the data from the EFU "consisted of Matching Technicians (Techs) and Matching Review Specialists (MRS), the highest level and most trained of the matchers from the PES."

3.5 The Evaluation Strata

Because the sample sizes in the evaluations, particularly in the EFU, are small when allocated down to the 1,392 poststrata, most of the information in the evaluation reports is aggregated to the 13 evaluation strata defined in Table 4. In particular, these 13 strata are used in [P16], the total error report. These strata are defined by geographic region, central city versus non–central city and minority versus nonminority. We will categorize information either using these same 13 strata or by aggregating to the national level. Table 4 lists the raw DSE undercount estimates for each evaluation stratum [P16]. They have considerable variation, being large in the five minority strata (1, 3, 5, 8 and 11) and smaller in the nonminority strata. Secondary variations appear by region and place type.

4. EFFECTS OF SMALL PES ERRORS

The two foregoing section's have described the production of the PES undercount estimates and described the evaluation data and reports, and they form the backdrop for our error analysis. However, to assess the effects of errors in the PES we need to cross another bridge. It turns out that small errors in the PES can lead to large errors in the undercount estimates. There are two reasons for this.

4.1 The Equation for Computing Undercount

Suppose that in a certain population pool, the census count was N_{C}. If N_{EE} persons were erroneously enumerated, then $N_{\mathrm{CE}} = N_{\mathrm{C}} - N_{\mathrm{EE}}$ is the number of persons correctly enumerated in the pool. The persons enumerated by the census form the E-sample.

Later, another survey of the same pool enumerates N_{P} persons (the persons in the P-sample) and, of these, N_{M} can be matched to the E-sample. Then the dual system estimate [except for a technical adjustment in how N_{CE} is computed (Mulry and Spencer (1991)] for the total number of persons in the pool is defined by

$$N_{\mathrm{DSE}} = N_{\mathrm{CE}} \times (N_{\mathrm{P}}/N_{\mathrm{M}}).$$

The undercount estimate is $N_{\mathrm{DSE}} - N_{\mathrm{C}}$ and the percent undercount is this difference expressed as a percent of N_{DSE}. Suppose, for example, that $N_{\mathrm{C}} = 1{,}020$, but it is determined that $N_{\mathrm{EE}} = 10$ so N_{CE} is 1,010. Suppose also that the number of people found in the second survey is $N_{\mathrm{P}} = 1{,}000$, and that there are $N_{\mathrm{M}} = 970$ matches. Then $N_{\mathrm{DSE}} = 1{,}041$ and the undercount estimate is 2.0%.

All three numbers $N_{\mathrm{CE}}, N_{\mathrm{P}}$ and N_{M} are estimates subject to error. In particular, N_{P} is subject to some of the same errors as N_{C}. The undercount estimates are sensitive to errors in N_{P} and N_{M}, less so to errors in N_{EE}. For instance, if N_{P} decreases by 0.5% to 995, and N_{M} increases by 0.5% to 975, then N_{DSE} decreases to 1,031 and the undercount estimate to 1.0%. On the other hand, if the two changes go 0.5% in the other direction, then the undercount estimate increases to 3.1%. Thus we have the following:

> Two 0.5% errors in estimating N_{M} and N_{P} can result in a 50% error in the undercount estimate.

4.2 The Weighting Effect

The effect of errors is further complicated by disparities in weighting. To get the DSE population estimate in a poststratum, the numbers used in the formula above are weighted up from numbers in the sample blocks. The 5,290 block clusters in the PES were not randomly selected from all U.S. block clusters. Instead they were randomly selected from predefined sampling strata.

On the average, one person in the PES corresponds to 650 in the U.S. population, but because of the stratification and nonresponse this can be uneven. There are some block clusters where one person weights up

TABLE 4
Evaluation strata and their estimated undercount

Location	Place type	Race–ethnicity	DSE undercount (%)
1 Northeast	Central city	Minority	6.83
2 Northeast	Central city	Nonminority	−0.75
3 United States	Non–central city	Minority	5.43
4 Northeast	Non–central city	Nonminority	0.01
5 South	Central city	Minority	5.68
6 South	Central city	Nonminority	1.94
7 South	Non–central city	Nonminority	1.82
8 Midwest	Central city	Minority	3.97
9 Midwest	Central city	Nonminority	1.28
10 Midwest	Non–central city	Nonminority	0.39
11 West	Central city	Minority	6.14
12 West	Central city	Nonminority	2.13
13 West	Non–central city	Nonminority + Indian	1.84

to over 10,000 in the U.S. population. Mistakes in a handful of people in such block clusters would be highly magnified. For example, in a certain block cluster a single unmatched family consisting of five persons contributed 45,000 to the undercount estimate.

Another illustration consists of two PES block clusters in which a low match rate increased the national undercount estimate by almost one million people. There were 648 persons in the two block clusters. The problem is discussed in a 1991 Census Bureau memorandum (Hogan, 1991). The cause was investigated and determined as faulty census geocoding. The two block clusters had their influence downweighted so they contribute only 150,000 to the estimated undercount.

The question is not whether the PES was accurate compared to other sample surveys or whether the matching was accurate compared to other matching projects, but whether they were accurate in terms of the resulting undercount estimates. Small errors in estimating N_P and N_M lead to large errors in undercount estimates. Errors in a small number of persons can be disproportionately magnified by the weighting.

5. ERROR SOURCES AND DATA QUALITY

Estimates for some types of errors in the national and in the 13 evaluation strata undercount estimates can be based on the evaluation data. These are covered in the next section and show that well over half of the national DSE undercount estimate is due to bad data.

However, there is a considerable amount of other information in the P-studies that is relevant to data quality issues and potential error sources in the DSE estimation process. This information is diverse and comes from many different reports. Putting these different pieces together gives insight on the difficulty of the DSE estimation process. The information available covers the following:

- matching errors;
- fabrications;
- census day address errors;
- geocoding errors;
- unreliable interviews;
- missing data;
- imputation.

Because many different facets of the data quality issue are covered, this section contains a fair amount of detail. At the end, we summarize and look at the implications.

5.1 Matching

Matching records from two different files of human data with differing names, ages, missing sex or race identifiers and different addresses can involve difficult decisions. In the present situation, one file consisted of the PES records. The other file consisted of the census records. The evaluation material concerning matching comes from a rematching study, from ongoing quality assurance records and from the reinterviews. Error rates for matching weighted to the nation and the 13 evaluation strata were derived from the rematching study and are given in [P16].

In this section we examine disagreement rates as totaled over individual cases. The marginal disagreement rates can be much smaller. For instance, weighted to the total population, the number of P-sample matches in the rematching study differs from the number in production matching by only 0.18%, while the disagreement rate is 1.8%.

TABLE 5
P-sample match–rematch disagreement rates

Whole sample	1.8%
Unresolved group	23.8%

TABLE 6
E-sample match-rematch disagreement rates

Whole sample	2.1%
Unresolved group	35.4%

TABLE 7
SMG1–SMG2 *disagreement rates*

Matched	10.7%
Not matched	6.6%
Unresolved	31.2%

However, similarities in highly aggregated marginal totals cannot be used to infer that match–rematch differences are small at the level of the 1,392 poststrata. Substantial poststrata differences can "average out" to small differences in marginal totals at aggregated levels. For instance, in the 13 evaluation strata, on average the number of P-sample matches in the rematching differs from the number in the production matching by 0.41%. This is over double the 0.18% difference found using the aggregation to the total population. We can expect larger differences at the state and city level.

5.1.1 *Rematching study data.* The primary sources of information about matching errors are the rematching study reports [P7] (P-sample) and [P10] (E-sample). In the study, over 70,000 E- and P-sample persons were rematched by more experienced matching teams and compared with the original matching. One indication of matching accuracy is in the disagreement rates between the category assigned by the production matching and that assigned in the rematch.

The disagreement rates for the P-sample persons are obtained as follows. In the production matching, each person in the P-sample was categorized as a match, nonmatch, unresolved or out-of-scope. The production out-of-scope persons were assigned zero weight in the sample and their numbers do not appear in the tables. The rematching did a similar categorization. The total disagreement rate is the number of people categorized differently in the two matching procedures as a percentage of the total number categorized.

Tables 16–28 of report [P7] give cross-tabulations of counts in the production versus rematch categories in each evaluation stratum with numbers weighted to the total population. The disagreement rate was computed for each evaluation stratum and averaged over the strata to give the first row of Table 5.

The second row of Table 5 gives the disagreement rate between the match and rematch teams on the membership of the unresolved category. This number is computed as the percentage of all persons categorized as unresolved by the rematch team that are categorized differently in the production matching. These numbers are also weighted to the total population and averaged across evaluation strata. Cases in the E-sample get put into three categories: correctly enumerated; erroneously enumerated; and unresolved. The primary evidence concerning errors in this categorization comes from the rematching study report [P10]. Table 6 summarizes the extent of the disagreement calculated the same way as for Table 5 and based on [P7, Tables 42–54].

5.1.2 *Quality assurance results.* Another source of information concerning disagreement in production matching is given in report [P8]. In the production matching process, the first phase was computer matching. This matched 75% of the cases. After the computer matching and some clerical matching, two teams (SMG1 and SMG2) worked, independently of each other, on the cases not matched by the computer.

Report [P8, Table 3.1] gives data concerning the disagreement rate between these two teams. We give the results since they comprise the only data available where two teams worked independently on matching the same cases. The breakdown in Table 7 gives the percentage disagreement on the major match categories. Overall, the disagreement rate was about 10% on the cases handled by the two teams, that is, the cases not matched by the computer. The disagreement rates were computed taking the SMG1 results as the base, that is, 10.7% of the cases categorized as "Match" by the SMG1 team were put into other categories by the SMG2 team.

5.2 Fabrications in the P-Sample

From [P5a], "Interviewers may fabricate people in the P-sample housing units. The creation of fictitious individuals has the effect of decreasing the PES match rate causing the estimate of coverage error [undercount] to be too large." Also, the effect is differential. The general belief is that the more difficult the area is to survey, the higher the fabrication rate

(Stokes and Jones, 1989). Thus, one expects high fabrication rates in minority, central-city areas—exactly those areas which have the highest estimated undercounts.

There are three studies estimating the extent of fabrications in the PES. The first is the evaluation field study; the second uses the data from quality control; and the third is the [P6] project, which attempted to quantify fabrication rates by looking at interviewers with unusually high nonmatch rates. The estimates of error due to fabrications used in [P16] were based on the P-sample data in the EFU.

Out of 14,444 cases in the P-sample EFU data, 13 were identified as fabrications, including 12 blacks [P5a]. These 13 weight to the national total as 0.03% of the cases. Based on the [P16] estimates, this 0.03% rate inflated the DSE national undercount estimate by 50,000 persons (see Section 6.1, Table 15). Thus, at a similar scaling, an undetected fabrication rate of 1% would have inflated the DSE undercount estimate by 1,650,000.

The EFU fabrication estimate may be low. To quote from report [P5a], "The data for the study were collected in the EFU which was not designed specifically to detect fabrication.... Thus, it is possible that the EFU did not identify more cases as fictitious because there was not enough new and additional interviewer information to establish that the cases were fictitious in the PES."

Report [P5] gives estimates of the fabrication rates based on quality control data gathered during the PES. During operations, quality control (QC) found that 0.26% of the household interviews were fabricated. Their estimate is that 0.06% of the remaining cases on a national level are fabrications. It is difficult to know how firm the basis is for this estimate. In particular, report [P5, page 4] states: "A limitation for this project (estimation of the fabrication rates) is the incompleteness and inconsistency of the QC data sources across RCC/ROs" (where RO is regional offices, and RCC is census centers).

Report [P6] tries to estimate the fabrication rate by identifying interviewers with an unusually high nonmatch rates. Of these interviewers, only 38% were identified as problem interviewers by quality control procedures. The report [P6, page 12] states: "It has been the speculation that in data collections such as the Census Bureau's current surveys between $\frac{1}{2}$ and $1\frac{1}{2}$% of the interviews are fabricated (Biemer and Stokes, 1989). The results from this study indicate that in an undertaking such as the PES the percentage is higher. Here, with the exception of two regions, the range was from 2.1 to 5.97%. In two regions, the percentage went as high as 7.79 and 8.79."

The report concludes [P6, page 15] that "Overall, between .9 and 6.5% of the interviewers were found to have high nonmatch rates. This compares favorably with the expectation that between 2 to 5% of interviewers are dishonest in their data collection." The estimates in [P6] are based on fitting mathematical models and on assuming that high nonmatch rates for an interviewer as compared to interviewers in neighboring blocks is a strong indication of fabrication. There is no data available to verify these assumptions.

At present, the effects of P-sample fabrication on the DSE estimates of undercount are difficult to quantify. The QC and EFU estimates seem low for reasons given above. The [P6] estimates seem high. The potential range is large, going from 0.03% to 8.79%. On the E-sample side, fabrication is treated as a component of erroneous enumeration. Report [P9a], which uses the E-sample EFU data to estimate errors in the erroneously enumerated counts, does not break out E-sample fabrication separately.

5.3 Census Day Address Errors

One of the most difficult sources of error to pin down is errors in the location of residence on census day (April 1, 1990). People who moved into the sample blocks after census day would not have been enumerated by the census as living in the sample blocks. If they appear in the PES sample, the nonmatch rate is erroneously inflated.

The EFU study found 334 P-sample respondents who were originally classified as nonmovers but new information revealed as after census day inmovers. Weighted to a national level, this represents 1,410,000 persons [P4]. These cases were then rematched using the new information. The computations in report [P16] show that this resulted in a decrease of 811,000 people in the estimated national undercount.

In total (nationally weighted) 41% of the newly discovered movers were originally matched in the PES. The implication is that the matching process incorrectly matched people not resident in the area on census day to people counted in the area by the census on the census day. Although the absolute number of these people is small, they weight up to 580,313 nationally.

5.4 Geocoding Errors

Geocoding, in this context, is the assignment of housing unit addresses to the selected sample blocks. The census makes one such assignment and the PES another. Errors in geocoding affect undercount estimates. Suppose the PES erroneously assigned a housing unit to one of the sample blocks. Then the P-sample persons in the unit could probably not be matched against any E-sample persons, and the nonmatch rate would be inflated. These errors are the subject of [P11], and the following discussion is based on that report.

In an effort to minimize the effects of geocoding errors, search areas consisting of one or two rings of blocks surrounding a sample block were defined. If an E-sample case cannot be matched to a P-sample case in its block, then it goes to follow-up and the interviewer is instructed to draw a sketch of the location of the housing unit. This sketch is then used to get a geocoding. If the location is in the sample block or search area and the enumeration is correct in other respects, the case is classified as correctly enumerated (CE); if not, as erroneously enumerated (EE). For a P-sample case in the block with no match in the block, a match in the search area is sought. If one exists, the case is put into match status.

Overall, 4.08% of the P-sample was matched to the Census through geocoding to the surrounding blocks. However, only 2.29% of the E-sample got CE status in surrounding blocks. The difference, weighted to the national level, is "an approximate excess of 4,296,000 in the P-sample population" [P11, Attachment]. The implication of this result is that if the surrounding blocks search had not been done, then geocoding errors would have caused a doubling of the DSE national estimated undercount, to over 4%. On the other hand, using a larger search area might well have produced a much lower undercount estimate.

The EFU included some of the E-sample households that had sketches made of their locations and geocoded in the production follow-up. For these households, the EFU interviewers made a second location sketch and a second geocoding was performed. Putting the geocodings into three categories (located inside a sample block, located in a search area, located outside of both) there is a 20% disagreement rate between the two geocodings [P11, Table 3.1].

5.5 Reliability of Interview Data

Some information reflecting PES interview reliability can be obtained from the EFU. Of the EFU P-sample interviews 4% were rejected as being unreliable [P4]. The EFU interviewers were regular census employees and more experienced than the PES interviewers. Thus, the 4% rejection rate seems surprisingly high. Of the rejected interviews 58% were with family members, 13% with neighbors, 13% with the apartment manager or landlord and 16% other.

The rejection rate is higher for minorities and central cities. Report [P4, Table 5.1.4] gives the ratio of the number of rejected interviews to the total number of EFU interviews broken down by evaluation strata. It is generally large where the estimated undercount is large. An analysis of the implications of this 4% rejection rate on the accuracy of the DSE estimates has not been carried out.

TABLE 8
Percentage change in match status using new EFU information

Correctly enumerated	7.2%
Erroneously enumerated	32.8%

The EFU collected data for 11,992 E-sample persons [P9a]. The new interview information was given to the matching team along with the PES production matching information. Match status could be changed from the PES production match status only if new, relevant and reliable information regarding a case was present in the EFU interview. The cross-tabulated data, weighted to the nation, comparing status before and after use of the EFU information is given in [P9a, Table 35]. Changes are summarized in Table 8.

Over 2,000,000 persons classified as "correctly enumerated" in the PES became classified either as "erroneously enumerated" or "unresolved" after use of EFU data. Over 1,600,000 persons originally in the "erroneously enumerated" category moved to "correctly enumerated" or "unresolved." The implication is that a substantial fraction of the interview data did not give reliable results in the original PES matching. The analogous data for P-sample persons does not appear in any of the P-reports.

5.6 Missing Data

Interviews can result in missing information for some of the people in the household. Missing data can affect the PES estimates in two ways. First, it can make matching more difficult and error-prone. Second, assignment of persons to a poststratum depends on some questionnaire characteristics. If these are missing, the person may be assigned to the wrong poststratum.

Report [P2] contains relevant data, weighted to the nation. Table 9, based on [P2, Table 3.3] gives the percentage missing for some PES and census questionnaire characteristics. The percentage of missing data is highest in those strata where the estimated undercount is highest. This property is true not only for race, but for all other characteristics. Table 10 gives the correlations between percentage of missing characteristics and percentage undercount over the 13 evaluation strata [P2, Table 3.4].

TABLE 9
Percentage of missing data

Characteristic	P-sample	E-sample
Race*	2.5	11.8
Age	0.7	2.4
Sex	0.5	1.0
Tenure	2.3	2.5

*Report [P2] states "the race variable . . . is a combination of race and Hispanic origin."

TABLE 10
Correlations of undercount estimates with percentage of missing data

Characteristic	P-sample	E-sample
Tenure	0.5	0.8
Sex	0.5	0.8
Age	0.6	0.7
Race	0.7	0.8

After matching, the missing data is filled in by a hot-deck imputation algorithm (Diffendal and Belin, 1991, Appendix 2). This serves two purposes. One is to allocate the persons to poststrata. The second is that complete information is necessary for the imputation of the unresolved persons into match categories. This latter procedure is discussed in the next section.

5.7 Imputation of the Unresolved Cases

At the end of production matching, there were 5,359 unresolved persons in the E-sample, and 7,156 unresolved persons in the P-sample [P1]. These 12,515 are among the persons having the most incomplete and least reliable data in the PES and the census. Over two-thirds of the unresolved people in the PES sample are after census day movers (52%) or possible movers (16%), and 45% are minority (Diffendal and Belin, 1991, Table 3.6). In the E-sample 32% have unresolved geocoding [P1].

Although the unresolved account for only 1.6% of the total combined PES and census samples, the estimates of the undercount strongly depend on what category they are finally assigned. If all unresolved PES sample cases are assumed to be matches and all census sample unresolved assumed to be erroneously counted by the census, then the DSE national estimate is 1,000,000 less than the census. At the other extreme, the DSE estimate is 9,000,000 more than the census.

The Census Bureau handles the unresolved P-sample by using a complex hierarchical logistic regression model that depends on estimating coefficients for dozens of variables (for details see Belin et al., 1993, and Diffendal and Belin, 1991). For each unresolved person in the P-sample, the P-sample model is used to compute a match probability. In each poststratum, the number of matches is increased by the sum of the match probabilities of the P-sample unresolved persons in the stratum.

The E-sample unresolved are treated using a different hierarchical logistic regression model (see Belin et al., 1993, and Diffendal and Belin, 1991). For each unresolved person in the E-sample, the E-sample model is used to compute a correct enumeration probability. Then the number of correctly enumerated people in each poststratum is increased by the sum of the correct enumeration probabilities of the E-sample unresolved people in the stratum. Neither model has been tested prior to their use on the 1990 PES data.

The outputs of these models have a significant and differential effect on the undercount estimates. In the five minority evaluation strata, imputation adds, on the average, 1.2% to the undercount estimates. For instance, in evaluation stratum 8, the imputation increases the undercount estimate from 2.8% to 4.0%, and in stratum 11, from 4.2% to 6.1%. In the nonminority evaluation strata, imputation adds an average of 0.6% to the undercount estimates. Thus, the imputation modeling is a significant contributor to the larger estimated undercounts in the minority strata. (These numbers were computed from Ericksen, Estrada, Tukey and Wolter, 1991, Table 11, Appendix C).

Significant proportions of key variables used in the models (such as age, sex, race, etc.) have been previously imputed. Of the P-sample unresolveds, 28% have at least one characteristic missing in their data. In the E-sample the percentage is 38% (Diffendal and Belin, 1991, pages 13 and 26). The coefficients of the variables are estimated using the data from the PES and the PES follow-up. This involves the further assumption that the final unresolved group is similar in nature to the persons resolved in the PES follow-up.

The only data available for assessment of the models comes from the EFU. After the imputation models were used to assign match and correct enumeration probabilities to unresolved PES persons, a subsample of these people were reinterviewed in the EFU. The new information was sent back to the matching teams and rematching was carried out. The results, weighted to the nation, are given in Table 11 (from [P3, Tables 3.1 and 3.2]).

The imputation models predicted that 42% of all of the P-sample persons in Table 11 would be matches and that 78% of all of the E-sample persons would be correctly enumerated [P3, Tables 3.3 and 3.4, nationally weighted]. The large proportion of the cases

TABLE 11
Rematch results for unresolved groups

P-sample unresolved		E-sample unresolved	
Match	12%	Correct enumeration	62%
Nonmatch	27%	Erroneous enumeration	17%
Unresolved	59%	Unresolved	21%
Out-of-scope	3%		

left unresolved, particularly in the P-sample, makes conclusions uncertain.

In computing and using the match probability, the assumption is that a high computed probability of a match implies that the person is very likely a true match. Thus, one would expect that, for a P-sample unresolved person with a high computed match probability, additional information would show that the person is indeed a match. This can be examined by looking at the new match status (Table 12) resulting from the EFU reinterview information for different ranges of computed match probabilities (from [P3, Table 3.1]).

As the match probabilities increase, the proportion of resolved cases that result in matches increases, but so does the proportion of unresolved cases. Report [P3], in summary of Table 11, states: "Thus, for P-sample persons, the imputation process is consistent with EFU results. However, the high percentage of unresolved persons in the EFU (58.55 percent) may limit the utility of this result."

Table 13 for the new match status for the E-sample unresolved versus the imputed correct enumeration probability is taken from [P3, Table 3.2]. There is no evidence here of an association between the probabilities computed by the imputation model and the enumeration status as determined by the EFU reinterview information.

The most important thing about estimating the undercount is not its total magnitude, but its differential effect on the 50 states and on thousands of counties and cities. These differential effects are estimated using the allocation of each PES-surveyed person into one of 1,392 poststratum. On the average, there are 270 persons and 9 unresolved cases per poststratum. How these nine cases are resolved is an important determinant of whether there will be a high or low estimated undercount for that poststratum.

Evidence for the accuracy of the imputation models at more disaggregated levels is not available. Report [P3] does not give a table of the imputation results by evaluation strata. In the total error report [P16], the imputation estimates in each of the 13 evaluation strata are assigned zero bias. No explanation is given.

Other ways of looking at the imputation results are presented in Belin et al. (1993). If attention is confined to the 316 P-sample persons resolved in the EFU (first two numbers of first column, Table 11), the model gives an accurate prediction of the proportion of matches (Belin et al., 1993, Table 3, page 1157). The predicted proportion of matches categorized by imputed match probabilities (as in Table 12, first two rows) is not as accurate. The results in Belin et al. (1993), are not comparable to Tables 11 and 12 because weighting to the nation is not used and the categorization is different.

There is another evaluation report that deals with the imputation models ([P1], "Analysis of reasonable alternatives"). This work is intended as a sensitivity analysis and not as an assessment of accuracy.

5.8 Summary of Data Quality Evidence

It was noted in Section 4 that several errors of the size of $\frac{1}{2}$% could have a large effect in the national undercount estimates. The analysis of the evaluation data not only shows sources of error potentially larger than $\frac{1}{2}$%, but also many such sources, the following in particular:

1. In the PES, 6.3% of the interviews were with other than household members. In the PES follow-up, 19.2% were with other than family members.
2. The percentage disagreements in the P-sample rematching study are all well above 1%, and the average over the evaluation strata is 1.8%. In the E-sample, the match code disagreement averaged over the evaluation strata is 2.1%. In the E- and P-sample, the disagreement on the makeup of the unresolved groups averaged 27.6%.
3. The fact that 4% of the EFU interviews were rejected as being unreliable is disturbing, since the EFU interviewers were more experienced than the PES interviewers. The implication of 4% unreliable information in the EFU needs to be considered in judging the reliability of the PES data.
4. Substantial changes (7 and 33%) in enumeration status assigned in production matching resulted when a rematching was done using the EFU E-sample reinterview data. This is a reflection of the reliability of the interview data used in production matching.
5. The imputation models used to assign unresolved persons into match or nonmatch, correctly enumerated or not, are previously untested, and the EFU evidence concerning their performance is inconclusive. There is no evidence concerning

TABLE 12
EFU rematch results versus imputed match probabilities

	Imputed match probability			
New match status	0–25%	25–50%	50–75%	75–100%
Match (%)	6	13	17	15
Nonmatch (%)	44	26	15	7
Unresolved (%)	47	57	66	77
Out-of-scope (%)	3	4	2	1

TABLE 13
EFU rematch results versus imputed correct enumeration probabilities

	Imputed correct enumeration probability		
New match status	0–50%	50–75%	75–100%
Correct enumeration (%)	67	50	65
Erroneous enumeration (%)	10	26	15
Unresolved (%)	23	24	20

TABLE 14
Correlations with DSE undercount estimates

Percent unresolved P-sample	0.7
Percent unresolved E-sample	0.8
P plus E match–rematch disagreements	0.6
Missing data (average)	0.7
Rejected EFU interviews (P plus E)	0.5

accuracy at the poststratum level, or even at the level of the 13 evaluation strata.

There are other indications of serious errors: the EFU found 334 persons that moved into the area after census day but were not identified as inmovers by the PES. This number weights up nationally to 1,410,000 movers not correctly identified by the PES. The EFU had a 20% geocoding disagreement rate with the PES follow-up. Reported P-sample fabrication rates may be significant underestimates.

Correlations between the undercount estimates and data quality indicators computed over the 13 evaluation strata are shown in Table 14.

The correlations of the undercount estimates with measures of bad data indicate that it is difficult to gauge what is being measured by the undercount estimates. The data quality is worst where the undercount estimates are the highest—in the minority strata. One conjecture may be that the correlations of estimated undercounts with bad-data measures show that where there are large amounts of bad data there are also large real undercounts. However, the difficulty is that we do not know how much of the estimate is due to bad data. Section 6 gives evidence on this issue.

TABLE 15
Decreases in the DSE undercount estimates due to the evaluation data

	Number	Reason
	553,000	P-sample rematching*
	811,000	Census day address errors*
	50,000	Fabrications*
	624,000	E-sample rematching*
	−473,000	E-sample reinterview*
	290,000	Ratio estimator bias*
	183,000	Late late census data
	164,000	New out-of-scopes in rematch
	358,000	New out-of-scopes in reinterview
	537,000	P-sample reinterview
	128,000	Reinterview of noninterviews
	1,018,000	Computer coding error
Total	4,243,000	

6. CHANGES IN THE DSE ESTIMATES INDICATED BY EVALUATION DATA

6.1 Summary of Decreases in Undercount

The original national DSE undercount estimate is 2.1%, or 5,275,000 persons. All through the evaluations, as more experienced personnel were used to collect data or to rematch, it was seen that the original DSE undercount estimates were too high. Report [P16] attempts to give an overview and summary of all DSE errors as estimated by the evaluation studies. In the last two years, studies of the DSE estimate errors have been published (Hogan, 1990; Mulry and Spencer, 1991, 1993). These are based, essentially, on the [P16] analysis.

There are omissions in [P16], and Table 15 gives a more comprehensive summary using both the applicable parts of the [P16] report and data gathered from the other P-projects as part of this report. The numbers given are the *decreases* in the DSE national undercount estimate indicated by the evaluation data, and their sources will be discussed below.

As a result of these decreases, the corrected undercount estimate is 1,032,000, or 0.4%. The corrections reduce the estimated undercount to about one-fifth of the original DSE value.

The first six entries in Table 15, marked with an asterisk, are listed in report [P16] and are computed from the data in Tables 1–13 of [P16]. Each of these tables lists, for each evaluation stratum, the numbers (in the column labeled "mean") that get substituted into the formula at the bottom of page 5 of [P16] to correct the DSE estimate. This was done, one error at a time, in the order in which they were listed in the [P16] tables. The results were then added across evaluation strata to give the tabulated results.

Report [P16] gives a second set of undercount estimates that includes an additional error source called model bias, more commonly known as correlation bias. This is the bias ascribed to the existence of persons unreachable by any survey. However, because these bias estimates are (and must be) based on highly speculative assumptions and have only a tenuous connection with any data, they are not included in our discussion.

The results of this paper were circulated to the Census Bureau in March, 1992. In the reviews there was agreement with the numbers given in Table 15 except, perhaps, in two areas. The first is the question of what is included in "census day address error." The second is the possibility that there are compensating factors to the "new out-of-scope" errors. The issues not resolved with the Census Bureau are discussed in the Appendix.

Using the bias estimate in the original version of [P16] drops the undercount estimate from 2.1% down to 1.4%. A later (June 1992) total error report (Mulry, 1992b) using different evaluation strata states the DSE as 1.6% with an upward bias of 0.7%, leading to a bias corrected estimate of 0.9%. Thus, the Census Bureau has come about halfway toward the estimate given in this paper. Using their current estimates, 55% of the original DSE undercount is due to bad data. Our estimate is 80%.

6.2 Error Sources Not Included in [P16]

There are six sources of error listed above which are not included in the [P16] report.

6.2.1 *Late late census data.* Some census data came in after the DSE estimates were computed. Because of time constraints, a compromise procedure was worked out which used only part of the late census data. In the blocks most affected, the late census data was matched to the P-sample data and the DSE estimates revised accordingly. Report [P18] estimates that if all of the late census data had been used, the DSE undercount estimate would be reduced by 183,000 people. The relevant descriptions and tables are contained in report [P18].

6.2.2 *New out-of-scopes.* P-sample out-of-scopes are cases that do not belong in the P-sample. They are out-of-scope for reasons such as being duplicate records, fictitious records, wrong addresses and so on, and they should be subtracted out of the number of persons in the P-sample.

Both in the P-sample rematch study and in the evaluation P-sample reinterviews, many cases that were originally classified as nonmatches were reclassified as out-of-scope. The estimated decrease in the undercount was obtained by decreasing the size of the weighted P-sample by the weighted number of persons switched from nonmatch to out-of-scope, and then recomputing the DSE estimate.

The out-of-scope P-sample corrections come from two sources. The data from the rematch study is taken from report [P7]. The reinterview data does not appear in the evaluation reports and was supplied by the Census Bureau upon request. To avoid overlap with census day address error analysis, the reinterview data used does not include the inmovers reported on in [P4].

6.2.3 *P-sample reinterview.* In the evaluation, the EFU P-sample reinterview persons were rematched using the new information. The changes in match status were given in tables similar to the tables in the P-sample rematch study. The number of new matches in each evaluation stratum was computed using the same method as the Census Bureau used to compute the number of new matches from the tables in the P-sample rematch report [P7]. The data used was supplied by the Census Bureau and does not include inmovers.

6.2.4 *Noninterview error.* Some households that were treated by the PES as noninterviews (refusals, no one home, etc.) and adjusted for by being "weighted out" were revisited in the EFU. The EFU was successful in getting interviews from 75% of these households. This new data was then supplied to the matching teams. As a result, report [P3] states: "At the national level, an estimated 102,403 more matches than are indirectly added by the noninterview adjustment would be added to the PES total." The relevant data is in report [P3], which gives, by evaluation stratum, the increased number of matches.

6.2.5 *Computer coding error.* In late 1991 the Census Bureau found an error in its computer code which resulted in sometimes classifying E-sample persons as correctly enumerated when they should have been classified as erroneously enumerated (Hogan, 1993). Correcting this error gave a revised undercount estimate slightly over 1,000,000 lower than the original estimate. The data on the effects of the coding error by evaluation stratum were transmitted to me in February 1992 by the Census Bureau. The numbers given in the June 1992 error report (Mulry, 1992b) indicate that only a minor portion of the error overlaps with other sources of error treated in [P16].

6.3 Decreases by Evaluation Stratum

Table 16 gives the distribution of the undercount changes by evaluation strata. The first column is the original DSE estimate undercount estimate. The second column gives the undercount estimates corrected as shown in [P16]. The third column gives the estimates corrected for the additional evaluation data noted above, and the fourth column gives the total change downward.

One trend seen in Table 16 is that the largest decreases occur in those strata having the highest original estimated undercount. In the five minority strata, the average decrease is 2.7%. In the nonminority strata, the average decrease is 1.3%.

6.4 Potential for Further Decreases and Changes

Even though the evaluation evidence cuts the undercount estimate by 80%, the available figures are probably conservative. The dependence in the rematch study tends to minimize discrepancies. The real fabrication rate may be larger than that used [P16]. The effect of the imputation models is largely unknown. Besides these, there are other potential changes not previously discussed and not included in Table 15.

(1) In the P-sample rematching and reinterview, 413,000 persons switched from an original classification of unresolved to out-of-scope. Many of these persons were imputed as nonmatches in the DSE estimates and should be treated as switching from nonmatch to out-of-scope. Using the best available estimates of the numbers of imputed nonmatches among these switchers lowers the undercount estimate by an additional 210,000. Most of the decrease affects the five minority evaluation strata, with their average undercount estimate going from 2.9% to 2.7%. The average undercount estimate in the nonminority strata stays about the same.

(2) The final undercount estimate for evaluation stratum 12 seems odd. Even though it is a nonminority stratum, it winds up having the third highest estimated undercount (3.1%). This is higher than three of the minority strata and is over twice as large as any other nonminority stratum. This may be due to a mistake in the census computation (see the Appendix). If so, the undercount estimate is further reduced by 236,000 persons, and the final estimate in stratum 12 drops to 1.1%. This latter figure is consistent with the other nonminority strata.

(3) Another problematic area is the fact that with a search area of 1–2 rings of blocks around a selected block, 4.1% of the P-sample were matched in the surrounding blocks, while only 2.3% of the E-sample got correct enumeration status in surrounding blocks. Suppose the Census Bureau had decided to use a search area of 6–8 rings of blocks. With this much larger search area more matches and more correct enumerations would be found in the surrounding blocks.

To see what the magnitude of the change might be, say that both rates increased by 20%, that is, the number matched in the surrounding blocks went to 4.9% of the total number of matches instead of 4.1%, and the correct enumerations to 2.8% instead of 2.3%. These assumptions seem conservative, but the undercount estimate would decrease by another 1,000,000.

There is some evidence that shows that the estimated undercount would significantly decrease with a larger search area. Concerning the two block clusters mentioned in the introduction that contributed almost one million to the estimated undercount, the June 1991 Census Bureau memorandum states: "the matching ... had been done correctly. However, approximately 75 percent of the non-matching people could have been converted to a match if the search area had been expanded."

Another piece of evidence is that in the rehearsal for the PES in Los Angeles in 1986, although relatively few households were matched outside their

TABLE 16
Undercount changes by evaluation stratum

Stratum	Original DSE	[P16] corrected	Total corrected	Change
1	6.8	5.4	3.7	3.2
2	−0.8	−1.3	−2.5	1.8
3	5.4	3.9	2.6	2.9
4	0.0	−1.1	−1.6	1.6
5	5.7	3.2	1.5	4.1
6	1.9	1.1	0.3	1.6
7	1.8	1.6	0.6	1.2
8	4.0	4.2	3.0	1.0
9	1.3	0.5	−0.5	1.8
10	0.4	−0.1	−0.7	1.1
11	6.1	5.7	3.9	2.3
12	2.1	3.7*	3.1*	−1.0*
13	1.8	0.7	−0.2	2.1

* These numbers may be erroneous due to a possible mistake in report [P16]. See the discussion in the Appendix. Correcting the mistake gives the numbers 1.7, 1.1 and 1.0.

blocks, 38% of those that were matched outside were matched more than five blocks away (Wolter, 1987a).

(4) The Census Bureau carefully rematched 104 block clusters having large numbers of nonmatches and erroneously enumerated persons (Hogan, 1993). The result was a further decrease of 250,000 in the estimated undercount. This has not been included in Table 15 because the possible overlap with other error sources listed. Getting a decrease of 250,000 in the estimated undercount by rematching 104 out of 5,290 block clusters raises the question of what additional changes might result if the Census Bureau had the resources for similarly rematching the rest.

7. COMMENTS ON REFERENCES

References that describe the methods used in the adjustment rehearsals and the evaluations of the outcomes are as follows: Anolik (1988) for Mississippi 1986; Hogan and Wolter (1988), Schenker (1988) and Stokes and Jones (1989) for Los Angeles 1986; Anolik (1989) for North Dakota 1987; Childers and Hogan (1989a, b; 1990), Diffendal (1988) and Mulry and Dajani (1989) for Missouri and Washington 1988. These publications are interesting in that they describe the outcomes of early efforts that led to the methodology used in 1990. Because the rehearsals were smaller in scale than the 1990 effort, their evaluations were sometimes more detailed.

Lessons learned in the rehearsals are summarized in Hogan (1989). Discussion and planning for the 1990 adjustment are given in Anolik (1990), Biemer and Stokes (1989), Childers et al. (1987) and West, Corby and Van Nest (1989). The most comprehensive view of the problems to be faced in 1990 is in the Bureau of the Census document Wolter (1987a).

The Secretary of Commerce's decision (Department of Commerce, 1991b) came with a six-inch-high stack of back-up material and contains good summary descriptions of the census, the adjustment procedure and the evaluations, as well as other useful material. Differing views toward adjustment are contained. In particular, two committee reports have informative views of the outcome of the adjustment project. The Undercount Steering Committee consisted of Census Bureau statisticians, with a majority favoring adjustment (Undercount Steering Committee, 1991). The Special Advisory Panel consisted of statisticians from outside the Census Bureau and split evenly on adjustment. Because of this split, the Special Advisory Committee submitted a number of reports, with Ericksen, Estrada, Tukey and Wolter (1991) and Wachter (1991) being the most substantive. Another interesting and informed view appears in Freedman (1991) (also contained in Department of Commerce, 1991b).

APPENDIX: ISSUES UNRESOLVED WITH THE CENSUS BUREAU

After review of this paper by the Census Bureau, two issues remained unresolved. First, the possibility was raised that some of the persons originally classified as out-of-scope by the PES, with a subsequent reduction in P-sample size, might truly be in-scope. If so, then these should be added back into the P-sample size, therefore increasing the undercount estimate and canceling out some of the effect of the new out-of-scope errors.

The only data available in the P-studies concerning the original out-of-scopes is in a sample of 193

PES out-of-scope persons sent to the EFU. The results are given in Report [P3, Table 3.6]. Many of these did come back into scope, but then were usually classified as matches. The Census Bureau provided a weighting to the national level of Table 3.6. Based on these numbers and assuming a match rate of 90% and that 50% of the unresolved are imputed as matches, the overall effect would be to increase the undercount estimate by less than 12,000 persons.

The second problem came up when investigating the estimated undercount in evaluation stratum 12. The estimate is increased by an upward adjustment of 1.27% attributed to census day address error [P16, Table 12]. However, the formula and example initially given to us by the Census Bureau for computation of the census day address error shows that the adjustment can only be negative, and it is negative in the other 12 evaluation strata.

In December 1991, I was informed that the formula was incorrect, that the tables in [P4] were wrong and that the census day address error computation in [P4] and [P16] included all of the errors found in the P-sample reinterviews regardless of whether they were census day address errors or not. This was surprising, since this error is consistently referred to both in [P4] and [P16] as census day address error. Furthermore, some of the tables in [P4] are reproduced in the published paper titled "Address reporting error in the 1990 post-enumeration study" (West, Mulry, Parmer and Petrik, 1991).

I have been unable to obtain from the bureau any more specific information regarding their method for computing census day address error. If the P-sample interview error is actually included in the census day address error, 537,000 persons would be subtracted from the total decrease of 4,243,000 persons detailed in Table 15. The result would be to lift the estimated undercount from 0.4% to 0.6%.

ACKNOWLEDGMENTS

The learning and study process leading to this paper was helped by the cooperation of many persons at the Census Bureau who patiently went over complex details of the PES, matching and evaluations with me. John Thompson, Chief, Statistical Support Division, Bureau of the Census, provided invaluable assistance by pointing me toward the right people and setting up meetings with them.

David Freedman provided many helpful comments and discussion which resulted in a much improved paper and I am glad (more or less) that he got me involved in these census issues. Ken Wachter also contributed insightful comments and discussion, Mark Hansen carried out spotlessly clean computer runs and Robert Koyak provided important information. Terry Speed has been untiring in his insistence on clear prose and good organization; Robert Kass made suggestions which improved readability; and numbers of anonymous referees contributed to a better paper.

P-Project Reports

All of the P-project reports referenced are in one of the 1990 Coverage Studies and Evaluation Memorandum Series issued by the Statistical Support Division, Bureau of the Census, and are all dated in July 1991. They have the main title "1990 Post-Enumeration Survey Evaluation Project P(#)" followed by title and author. We list project number, memorandum series identification, title and author(s) for each of the reports referred to in the text.

P1 Series #A-9, Analysis of reasonable alternatives, Stephen Mack, Eric Schindler and Joe Schafer.

P2 Series #B-4, Distribution of missing data rates, Philip M. Gbur.

P3 Series #C-2, Evaluation of imputation methodology for unresolved match status cases, Philip M. Gbur.

P4 Series #D-2, Quality of reported census day address, Kirsten K. West.

P5 Series #E-4, Analysis of PES P-sample fabrications from PES quality control data, Antoinette Tremblay.

P5a Series #F-1, Analysis of P-sample fabrication from evaluation follow or -up data, Kirsten K. West.

P6 Series #G-2, Fabrication in the P-sample: interviewer effect, Kirsten K. West.

P7 Series #H-2, Estimates of P-sample clerical matching error from a rematching evaluation, Mary C. Davis and Paul Biemer.

P8 Series #I-2, Matching error—estimates of clerical error from quality assurance results, Michael Ringwelski.

P9a Series #K-2, Accurate measurement of census erroneous enumerations, Kirsten K. West.

P10 Series #L-2, Measurement of the census erroneous enumerations—clerical error made in the assignment of enumeration status, Mary C. Davis and Paul Biemer.

P11 Series #M-2, Balancing error evaluation, Randall Parmer.

P13 Series #O-3, Use of alternative dual system estimators to measure correlation bias, William Bell.

P16 Series #R-6, Total error in PES estimates by evaluation post strata, Mary H. Mulry.

P18 Series #T-1, The evaluation of late late census data in the 1990 post enumeration study, Nicholas Alberti.