Two Papers on Census Undercount Adjustment:

- **Straining Out Gnats and Swallowing Camels: The Perils of Adjusting for Census Undercount**

- **Quantifying Measurement Error and Bias in the 1990 Undercount Estimates**

Kenneth Darga, Senior Demographer
Office of the State Demographer
Michigan Information Center
Michigan Department of Management and Budget

April 29, 1998

Please address comments and suggestions regarding this analysis to:
DargaK@state.mi.us

EXECUTIVE SUMMARY

"Straining Out Gnats and Swallowing Camels: The Perils of Adjusting for Census Undercount"
" Quantifying Measurement Error and Bias in the 1990 Undercount Estimates"

April 29, 1998

There is reason to believe that the proposed remedy for Census undercount would be far worse than the undercount problem itself.

The proposed method of adjusting for undercount involves conducting a sample survey to identify people who were missed by the Census and people who were counted twice or counted in the wrong location. In order to succeed, this survey has to secure participation by the people who were missed by the Census, and it has to be very accurate in matching individuals counted by the sample survey with individuals counted by the Census. Unfortunately, these are impossible tasks: there are too many people who do not want the government to know where they are, and there are too many obstacles to matching the results of the two surveys successfully.

The undercount adjustments that were developed by this method for the 1990 Census seemed plausible at first glance, but they were strongly affected by several types of error in classifying people as missed or not missed by the Census. The first paper ("Straining Out Gnats and Swallowing Camels: The Perils of Adjusting for Census Undercount") shows that the proposed approach makes high levels of error inevitable and that the resulting adjustments have indeed been seriously flawed. The second paper ("Quantifying Measurement Error and Bias in the 1990 Undercount Estimates") identifies and quantifies several specific types of error:

- survey matching error
- fabrication of interviews
- ambiguity or misreporting of usual residence
- geocoding errors
- unreliable interviews
- unresolvable cases.

Together, these papers show that many of the people who were missed by the Census were missed by the coverage survey as well, and that many of the people who were identified as missed by the Census actually do not seem to have been missed at all.

Thus, in addition to reflecting differences in actual undercount rates, the adjustments derived from the sample survey reflect differences in the rate of error in classifying people as undercounted. Applying such adjustment factors to the Census would decrease the accuracy of local population counts and of the many detailed tabulations that are relied upon by all levels of government and by myriad private users of demographic data. These errors would usually be small, but they would sometimes be errors of 10%, 20%, or more. Since no one would know which areas and which population groups had serious errors, and since the errors would not be consistent from one Census to the next, all findings based on Census data and all comparisons between different time periods would come into question. In an attempt to address an inaccuracy at the national level, we would utterly destroy the reliability of Census data at the state and local level.

Straining Out Gnats and Swallowing Camels: The Perils of Adjusting for Census Undercount

Kenneth Darga, Senior Demographer
Office of the State Demographer
Michigan Information Center
Michigan Department of Management and Budget

April 29, 1998

Please address comments and suggestions regarding this analysis to:
DargaK@state.mi.us

Although the Department of Commerce is often criticized for Census undercount, it is not surprising that every Census misses a portion of the population. In fact, what is noteworthy is not that the undercount persists, but rather that the net undercount appears to have been less than 5 million people in 1990, or only about 1.8% of the population.[1]

A major reason for the undercount—although not by any means the only reason—is that quite a few people do not want their identities known by the government. For example, the United States has over 1 million people who do not make any of their required payments on court ordered child support[2] and an estimated 5 million illegal immigrants.[3] Each year, the police make over 14 million arrests for non-traffic offenses.[4] Millions of additional criminals remain at-large, many people would lose government benefits if the actual composition of their households were known, and many people have other reasons for concealing their identity and whereabouts from the government. If the Census misses fewer than 5 million people under these circumstances, then the Census Bureau is doing a truly remarkable job.

Nevertheless, eliminating even this small error would be a valuable achievement. Although the impact on many components of the population would be small, people in some demographic and economic categories are undercounted more than others. This leads to anomalies and imprecision in some analyses and affects political apportionment and fund distribution. The Census Bureau has therefore tried very hard to devise ways to measure and compensate for the problem of undercount.

Obviously, these methods are intended to make the Census count better. However, we need to evaluate their actual effects instead of their intended effects. Before we decide to use these particular methods in the official population count for the year 2000, we need to determine whether they would make that population count better or worse.

---

1 U.S. Department of Commerce, "Census Bureau Releases Refined 1990 Census Coverage Estimates from Demographic Analysis," Press Release of June 13, 1991, Table 1.

2 Economics and Statistics Administration, U.S. Department of Commerce, "Statistical Brief: Who Receives Child Support?," May 1995.

3 U.S. Immigration and Naturalization Service, "INS Releases Updated Estimates of U.S. Illegal Immigration," Press Release of February 2, 1997.

4 U.S. Department of Justice, Bureau of Justice Statistics, *Sourcebook of Criminal Justice Statistics*, 1995, p. 394.

After reviewing some of the reasons for believing that censuses miss a portion of the population, this paper briefly describes the Census Bureau's proposed method of adjusting for undercount. It will then be shown that, although the results of this method for 1990 appeared plausible, at least at the broadest national aggregation, the method cannot produce reliable adjustments for undercount: It is not capable of counting many of the people who are missed by the Census, it is very sensitive even to extremely small sources of error, and it is subject to many sources of error that are very serious. Thus, it is not surprising to find that many of the detailed undercount measurements for 1990 were implausible and, in some cases, demonstrably false. In an effort to correct a net national undercount of less than 2%, spurious undercounts of 10%, 20%, and even 30% were identified for some segments of the population. Adjustments derived from these measurements would have had a devastating impact on the usefulness and accuracy of Census data at the state and local level, and they would have had an adverse effect upon nearly all purposes for which Census data are used. Similar problems can be expected with the undercount adjustment proposed for Census 2000: The problems are not due to minor flaws in methodology or implementation, but rather to the impossibility of measuring undercount through the sort of coverage survey that has been proposed.

The Evidence of Undercount. Before examining the Census Bureau's method of adjusting for undercount, it is instructive to consider how we can know that each Census misses part of the population.

One way to find evidence of undercount is to project the population for a Census year by applying mortality rates and migration rates to the results of other censuses. The pattern of differences between these projections and the actual Census counts can provide good evidence for undercount. For example, if the count of black males age 20 to 24 is lower than would be expected based on the number of black males age 10 to 14 in the previous Census, and if it is lower than would be expected based on the number of black males age 30 to 34 in the following Census, then there is good evidence of undercount for that segment of the population.

The most widely accepted method for measuring Census undercount is called "demographic analysis." Using a combination of birth registration data, estimates of under-registration, mortality rates, estimates of international migration, social security enrollment data, and analyses of previous censuses, the Census Bureau develops estimates of the national population for each Census year by age, race, and sex. Although they are not perfect, the gap between these estimates and the national Census count provides the best available measure of undercount. The pattern of undercount suggested by demographic analysis is generally consistent from one Census to another, and it is consistent with the discrepancies that are found between

population projections and Census counts: Undercount rates appear to be higher for males than for females, higher for blacks than for whites, and higher for young adults than for people in other age groups.[5]

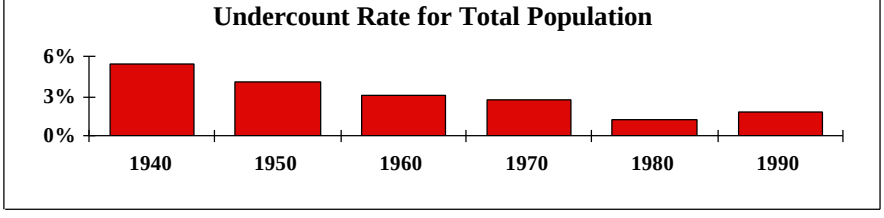
Demographic analysis suggests that the net national undercount fell in each successive Census from 5.4% of the population in 1940 to only 1.2% in 1980. This reflects improvements in Census-taking methodologies, special efforts focused on segments of the population that are hard to count, and assurances that Census information will be kept strictly confidential. However, the estimated net undercount rose to 1.8% in the 1990 Census: still quite low by historic standards, but disappointing because it represents an increase relative to the previous Census. (See Figure 1.)

A major shortcoming of this method is that it works only at the national level: There is too much interstate and intrastate migration to allow a phenomenon as subtle as Census undercount to be visible at the state or local level through demographic analysis. Since we can expect undercount to vary considerably from state to state and neighborhood to neighborhood, we cannot simply apply the national undercount rates to state and local population counts. This would not adjust some areas enough, and it would introduce inaccuracies into areas where there had not been inaccuracies before.

Figure 1

Estimates of Census Undercount Based on Demographic Analysis[6]

Population Category	1940	1950	1960	1970	1980	1990
Total Population	5.4 %	4.1 %	3.1 %	2.7 %	1.2 %	1.8 %
Black	8.4 %	7.5 %	6.6 %	6.5 %	4.5 %	5.7 %
Non-Black	5.0 %	3.8 %	2.7 %	2.2 %	0.8 %	1.3 %



[5] J. Gregory Robinson et. al., "Estimation of Population Coverage in the 1990 United States Census Based on Demographic Analysis," *Journal of the American Statistical Association,* 88(423):1061-1079.

[6] *Ibid.*, p. 1065.

Calculating Adjustments for Undercount. The Census Bureau has therefore tried to develop additional methods to estimate how well the Census covers each segment of the population. Immediately after the Census count is complete, the Bureau conducts a "coverage survey" which essentially repeats the population count for a small sample of census blocks. The coverage survey was called the "PES" or "Post-Enumeration Survey" in 1990, and it will be called "ICM" or the "Integrated Coverage Measurement Survey" in 2000. Data from the coverage survey are matched person-by-person with the original Census to identify the individuals counted by the coverage survey who seem to have been missed by the Census. These results are tabulated by relevant population characteristics to produce estimated undercount rates which can be applied to local areas based on their counts of persons with those characteristics. A sample of original Census forms are also matched with the coverage survey to identify individuals who were counted by the Census but omitted by the survey. These discrepancies are investigated and used to estimate "erroneous enumerations" or overcount.

Plausibility of the Adjustments. The resulting adjustment to the 1990 Census was quite plausible at the broadest national level. After moving up and down as corrections were made to the data and new statistical techniques were applied, the estimate of overall net undercount at the national level was 1.6%[7]—very close to the 1.8% suggested by demographic analysis. The credibility of the 1990 coverage survey was increased by the fact that it suggested high rates of undercount at the national level for the groups that would be expected to have high undercounts, such as Hispanics, blacks, people with difficulty speaking English, people in complex households, and people living in non-standard housing units.[8] Thus, one is tempted to conclude that the data from a coverage survey can provide an incredibly accurate measure of Census undercount.

Implausibility of the Adjustments. Before drawing that conclusion, however, we must consider a much less incredible interpretation: The differences between the coverage survey and the original Census may not represent net undercount as much as they represent the difficulty of matching individual records between two surveys. At a very broad level of aggregation, this methodological difficulty can produce

---

[7] Howard Hogan, "The 1990 Post-Enumeration Survey: Operations and Results," *Journal of the American Statistical Association*, 88(423):1047-1060, 1993.

[8] Manuel de la Puente, U.S. Bureau of the Census, "Why Are People Missed or Erroneously Included by the Census: A summary of Findings From Ethnographic Coverage Reports," report prepared for the Advisory Committee for the Design of the Year 2000 Census Meeting, March 5, 1993. J. Gregory Robinson and Edward L. Kobilarcik, U.S. Bureau of the Census, "Identifying Differential Undercounts at Local Geographic Levels: A Targeting Database Approach," paper presented at the Annual Meeting of the Population Association of America, April 1995.

results that look very much like net undercount because the population groups which are hard to match between surveys are generally the same groups that are hard to count. It is only by considering the tremendous barriers to measuring undercount accurately and by examining the detailed findings of the 1990 PES that we are led to accept this alternate interpretation. If this interpretation is correct, it has very clear implications for how the next Census should be conducted: Adjusting the new Census based on a coverage survey would negate the findings from 100 million Census forms based on a statistical artifact.

For a coverage survey to measure net undercount with anything approaching an acceptable level of accuracy, it must accomplish two impossible tasks. The impossibility of these tasks should lead us to question its validity even if it appears on the surface to provide a good measure of undercount. In particular, we should not conclude that the Census Bureau has accomplished the impossible merely on the basis of plausible results for the broadest national aggregation. If the detailed results do not make sense as well, then it is untenable to suggest that undercount has been measured with a high level of precision.

The first impossible task that a coverage survey must accomplish is to secure participation by two particularly problematic components of the population that are not counted well by the Census: homeless people and people who do not want to be counted. Each Census includes a major effort to count people in shelters and on the streets, but it undoubtedly misses a large portion of this population. A coverage survey is not well equipped to measure this component of the undercount because many homeless people are not likely to be found in the same place a few weeks or months later when the survey is conducted. The Census Bureau understands the impossibility of this task, and *the 1990 PES therefore did not even attempt to address this portion of the undercount.*[9] A coverage survey does not fare much better with the the other problematic component of the population. It is hard to imagine that very many of the people who avoided being counted by the Census are likely to be counted by a second survey that has essentially the same limitations. If drug dealers, fugitives, and illegal immigrants were afraid to fill out the Census form that everyone in the nation was supposed to receive, they are not likely step forward a few weeks or months later when their household is singled out for a visit by another government enumerator. On the contrary, they are likely to avoid the coverage survey even more studiously than they avoided the Census. Thus, we cannot believe that a coverage survey provides a good measure of undercount unless we are first willing to believe that somehow—without the tools necessary to do

[9] Howard Hogan, *op. cit.*

so—it manages to secure participation by these two groups of people who were not counted well by the Census.

If a coverage survey misses many of the same people who were missed by the Census, then the only way it can suggest a plausible level of undercount is by identifying other people as missed by the Census when they really were counted. This leads us to the second impossible task which a coverage survey must accomplish: achieving a practically-perfect replication and matching of Census results for that vast majority of the population which is counted correctly the first time. The problem is that, for every hundred people missed by a Census, there are about 3,000 people who were counted and can therefore be mistakenly identified as missed. These 3,000 people will inevitably include a certain number of challenging cases involving aliases, language barriers, individuals and households that have moved, people with no stable place of residence, and a host of other difficulties. It doesn't take a large error rate in classifying these 3,000 people who were correctly counted by the Census to completely invalidate our attempt to count the 100 people who were missed—especially since many of the people who were missed are making every effort to be missed again. A hypothetical example will help to demonstrate why even a 99% level of accuracy is not sufficient, and a review of the barriers faced by a coverage survey will demonstrate why 99% accuracy is not likely to be achieved.

Let's say that the next Census has an undercount of 3% and an overcount of 1%, for a net undercount of 2%. Let us also assume that the next coverage survey somehow manages to identify all of the people who are missed by the Census and all of the people who are counted twice or counted in error. This is a very generous assumption, since we have already seen that we have good reason to believe that this is an impossible task. Finally, let us assume that the coverage survey achieves 99% accuracy in classifying the individuals who were counted by the Census.

The apparent undercount will then include that 3% of the population which had been missed by the Census, plus nearly another 1% that had actually been counted correctly. This is because 1% of the 97% *not* missed by the Census will be falsely identified as undercounted because we achieve "only" 99% accuracy in replicating and matching the Census results. Thus, even under these unrealistically favorable assumptions, about 25% of the apparent undercount will actually represent classification error.[10] The measure of overcount will be even more problematic: It

[10] Expressed as a proportion of the actual population, the people counted by the Census who are mis-classified as uncounted in this hypothetical example will be (1.00 - .03) * (1.00 - .99) = .0097, where .03 is the assumed rate of undercount and .99 is the assumed level of accuracy. If we assume that all of the actual undercount will be detected through the coverage survey, the total estimate of undercount will be .03 + .0097 = .0397. Expressed as a proportion of the identified undercount, the people who are mis-classified as uncounted will therefore be .0097 / .0397 = .2443, or approximately 25%.

will include that 1% of the population that had actually been overcounted, plus nearly another 1% that had been counted correctly the first time. This means that about 50% of the apparent overcount will actually represent classification error.[11] This would hardly be a firm basis for fine-tuning the Census count.

Why the Word "American" Is Abbreviated in Census Questions

When you are trying to measure a small component of the population—such as people who have been missed by the Census—it is necessary to avoid even very small errors in classifying that vast majority of the population which is not part of the group being measured.

This principle is illustrated by one of the problems that the Census Bureau found while it was testing different ways of asking its new Hispanic-origin question for the 1980 Census. A very small number of people with no Mexican heritage thought that the category "Mexican or Mexican-American" meant "Mexican or American." Since they were "American," they thought that this category applied to them. Unfortunately, since people of Mexican heritage represented only about 4% of the national population, even this very small error among the remaining 96% of the population was enough to completely invalidate the count of Mexican-Americans. In fact, for many areas, a *majority* of the people selecting this category were found to be "Americans" with no Mexican heritage.

The 1980 Census therefore used the category "Mexican or Mexican-Amer." This was a big improvement, but the 1980 post-enumeration survey found that non-Mexicans still represented a majority of the people choosing this category in some areas with a very low population of Mexican-Americans. The 1990 Census therefore used the category "Mexican or Mexican-Am." This cleared up the problem.

A very similar difficulty arises when you try to measure undercount with a coverage survey. It is sometimes very hard to match up the people that you counted in the coverage survey with the people that you counted in the Census. When you make a mistake, people can be counted as missed by the Census or as mistakenly included in the Census when they really weren't. Since there are about 97 of these potential mistakes for every 3 people who were really missed by the Census, even a very low error rate is enough to completely invalidate the measure of undercount. Unfortunately, although the problem is very similar, the solution is not: Errors in matching surveys cannot be prevented by anything as simple as using more abbreviations.

[11] Expressed as a proportion of the actual population, the people counted by the Census who are mis-classified as counted in error will be (1.00 - .03) * (1.00 - .99) = .0097, where .03 is the assumed rate of undercount and .99 is the assumed level of accuracy. If we assume that all of the actual overcount will be detected through the coverage survey, the total estimate of overcount will be .01 + .0097 = .0197. Expressed as a proportion of the estimated overcount, the people who are mis-classified as counted in error will therefore be .0097 / .0197 = .4924, or approximately 50%.

A coverage survey must therefore achieve far more than 99% accuracy in classifying the people who are correctly counted by the Census. But is it possible to achieve such a high level of accuracy? Even for simple surveys conducted under ideal conditions, a 99% level of accuracy would be impressive. Unfortunately, the Census and the coverage survey are not simple, and they are not conducted under ideal conditions. The attempt to match the results of these two surveys must contend with a wide array of daunting problems, some of which are listed in the box on the following page. These problems are more than just hypothetical illustrations: many of them have been documented and quantified by analysts from the Census Bureau and elsewhere, who confirm that the undercount analysis involves very serious levels of matching error and other error. (See accompanying paper, "Quantifying Measurement Error and Bias in the 1990 Undercount Estimates.") Thus, in addition to knowing from logical arguments and hypothetical illustrations that serious problems are inevitable, we know from experience that serious problems actually do occur.

In place of our previous assumptions that a coverage survey measures overcount and undercount perfectly and that it matches the correct findings of the Census with 99% accuracy, we should therefore consider the implications of a somewhat more modest level of success. Let's say that the next coverage survey identifies 30% of the actual undercount and 40% of the actual overcount, that the undercount analysis averages an impressive 96.2% rate of accuracy in replicating and matching the correct results of the Census, and that the overcount analysis averages a similarly impressive 97.3% rate of accuracy. Although classification error would then account for an overwhelming 80% of the people identified as undercounted and 87% of the people identified as overcounted, the estimated net undercount at the national level would be the same 1.6% that was suggested by the coverage survey for 1990.[12] In other words, the estimate of undercount would primarily reflect errors in matching survey responses with Census responses, yet the broadest national estimate of net undercount would appear very plausible.

---

[12] Expressed as a proportion of the actual population, the the people counted by the Census who are mis-classified as uncounted in this hypothetical example will be (1.00 - .03) * (1.00 - .962) = .03686, where .03 is the assumed rate of undercount and .962 is the assumed level of accuracy. If we assume that 30% of the actual undercount will be detected through the coverage survey, the total estimate of undercount will be (.03 * .30) + .03686 = .04586. Expressed as a proportion of the identified undercount, the people who are mis-classified as uncounted will therefore be .03686 / .04586 = .8038, or approximately 80%.

The people counted by the Census who are mis-classified as counted in error will be (1.00 - .03) * (1.00 - .973) = .02619, and the total estimate of overcount will be (.01 * .40) + .02619 = .03019. Expressed as a proportion of the identified overcount, the people who are mis-classified as counted in error will therefore be .02619 / .03019 = .8675, or approximately 87%. The estimate of net undercount will be .04586 - .03019 = .01567 or 1.6%.

AN IMPOSSIBLE TASK

The Census Bureau tries to measure undercount by carefully taking a second survey for a sample of small geographic areas and comparing its results to the Census to see which persons had been missed. But is it possible to achieve a near-perfect match between these two surveys? This effort has to deal with daunting problems such as these:

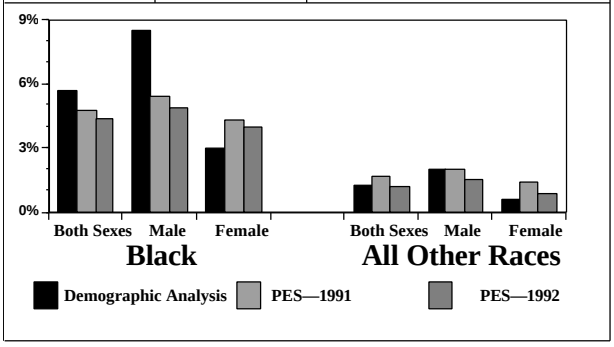
- Illegible handwriting.
- Similarity of names.
- Use of different nicknames and other variations on names in different surveys.
- Names which do not have a consistent spelling in the English alphabet.
- Use of aliases by illegal immigrants, fugitives, and others who place a very high value on privacy. Some people have more than one alias, some may use different names on different surveys, and some may be known to neighbors by names that are different from the ones used on the Census.
- Irregular living arrangements, complex households, and households with unstable membership.
- Differences which arise from collecting most Census information through written forms and collecting information for the coverage survey through personal interviews.
- Households and individuals that move between the Census and the coverage survey. (This is particularly a problem for college students, recent graduates from high school or college, and people who migrate between northern and southern states on a seasonal basis. Many of these people move within a few weeks after the April Census.)
- Differences which arise from having different household members provide information for the different surveys, or from having a responsible household member provide information for the Census and a child, neighbor, or landlord provide information for the coverage survey. (For example, differences in the reported name, age, race, or marital status can make it difficult to determine whether a person found by the coverage survey is really the same person found by the Census.) This problem was compounded in 1990 because the survey to measure undercount was centered around the Fourth of July weekend and the survey to measure "erroneous enumerations" was centered around the Thanksgiving weekend. It is very difficult, for example, to survey a college town during Thanksgiving week to determine who was living there the previous April.
- Language barriers. Language barriers are a particularly serious problem for a coverage survey because it relies upon personal interviews instead of on a written survey that respondents can complete with help from friends or other family members.
- People who are included on the Census but avoid inclusion on the coverage survey because they do not want to be identified by government authorities.
- Homeless or transient people who are enumerated in one housing unit by the Census but are in a different housing unit or on the streets at the time of the coverage survey.
- Homeless or transient people who are enumerated in the streets by the Census but are found in a housing unit by the coverage survey.
- Information that is fabricated by the enumerator or by the respondent.
- Clerical errors and processing errors.
- Failure to follow complex procedures precisely.
- Census forms which are coded to the wrong geographic area, making it impossible to to match them with the proper survey results.
- People who give an inaccurate response when they are asked where the members of their household were living on April Fools Day.

To people who are interested only in the national count of total population, the hypothetical example above may not appear very troubling. After all, since this example assumes that the errors in measuring undercount are largely offset by the errors in measuring overcount, the national population total it produces is actually closer to the assumed true population than the unadjusted Census count. What makes this example troubling is the fact that the undercount adjustments are relied upon for far more than a national population total. They purport to tell us which segments of the population and which parts of the country are undercounted more than others. The critical point that needs to be understood is that, if the coverage survey really does fail to measure a large portion of the undercount and if it mistakenly identifies people as missed by the Census who really weren't, then the differential undercounts it suggests will largely reflect differences in the amount of error in measuring undercount rather than differences in the amount of undercount itself. What would we expect such adjustments to look like? To put it simply, we would expect them to look just like adjustments developed from the 1990 Post-Enumeration Survey.

Figure 2

Alternate Estimates of Undercount for the 1990 Census[13]

Race and Sex	Basis of Undercount Estimate		
	Demographic Analysis	Post-Enumeration Survey	
		6/91 Revision	9/92 Revision
Black			
Both Sexes	5.7%	4.8%	4.4%
Male	8.5%	5.4%	4.9%
Female	3.0%	4.3%	4.0%
All Other Races			
Both Sexes	1.3%	1.7%	1.2%
Male	2.0%	2.0%	1.5%
Female	0.6%	1.4%	0.9%



[13] The undercount estimates based on the PES are from Barbara Everitt Bryant, "Census-Taking for a Litigious, Data Driven Society," *Chance: New Directions for Statistics and Computing*, Vol. 6, No. 3, 1993. The estimates based on demographic analysis are from U.S. Department of Commerce, "Census Bureau Releases Refined 1990 Census Coverage Estimates from Demographic Analysis," Press Release of June 13, 1991, Table 1.

At the national level, it would not be surprising for the undercount adjustments to look fairly reasonable: Since the population groups that are hard to match between two surveys are generally the same groups that are hard to count in the Census, we would expect the findings for very broad components of the population to be at least roughly similar to the results of demographic analysis. Of course they wouldn't be identical, since the level of difficulty in matching each group between surveys does not correspond precisely to the level of difficulty in counting it for the Census. For example, some problems such as language barriers and aliases pose more difficulty in survey-matching than in taking a Census, and segments of the population that are counted very well in the Census are at the greatest risk of having classification error exceed the actual level of undercount. Thus, while advocates of adjustment have not considered the pattern of differences displayed in Figure 2 to be unreasonable, the final national PES results for 1990 are actually quite different from the estimates based on demographic analysis even for very broad population groups. The apparent undercount for black males is 42% less than the rate suggested by demographic analysis, and the rate for white, Native American, and Asian/Pacific females is 50% higher. Under most circumstances, these differences would be considered very substantial.

We would expect an even worse situation below the national level. If the measure of net undercount is more sensitive to variations in the rate of classification error and other survey problems than to variations in the actual rate of undercount, it would not be surprising to find some serious deviations from the orderly pattern that would be found in a practically-perfect analysis. For example, it would not be surprising for the adjustment factors to look something like the ones displayed in Figure 3.

Figure 3 shows some of the initial undercount adjustments for children under age 10 which the Census Bureau developed based on the 1990 PES. This age group was chosen for this analysis because there is no obvious reason to expect householders to mis-report their young male children at a significantly different rate from their young female children. It is therefore disconcerting that these undercount adjustments for 1990 include some very large differences between boys and girls in this age group. In fact, these eighteen pairs of figures were selected for the table because they each have a discrepancy of *over ten percentage points.* It is even more disconcerting that these differences follow no discernible pattern. Sometimes the adjustment for boys is higher, but sometimes the adjustment for girls is higher; in one place black renters have a higher adjustment for boys, but in another place they have a higher adjustment for girls; in some places the gender discrepancy for whites is similar to the gender discrepancy for blacks, but in other places it is the opposite; sometimes one race category in a large city has a higher adjustment for boys, but

another race in the same city has a higher adjustment for girls. It is not surprising when signs of estimation error are visible for small components of the population in small geographic areas, but here we see apparently arbitrary adjustments for even the largest population groups in some of the largest cities and across entire regions.

Figure 3

Selected Undercount Adjustments for Children Under Age 10 from the 1990 Post-Enumeration Survey[14]

Region	Area Type	Tenure	Race	Adjustments	
				Male	Female
Pacific	Non-Central Cities	Renter/ Owner	Asian/Pacific	+ 5 %	+ 17 %
Mid Atlantic	Central Cities in New York City PMSA	Renter/ Owner	Asian/Pacific	+ 25 %	+ 9 %
East North Central	Central Cities in Metro Areas w/ Central City > 250K	Owner	Black	+ 26 %	+ 15 %
Pacific	Central Cities in Los Angeles PMSA	Owner	Black	+ 28 %	+ 8 %
Mid Atlantic	Central Cities in New York City PMSA	Owner	Black	+ 0 %	+ 23 %
South Atlantic	Central Cities in Metro Areas w/ Central City > 250K	Renter	Black	+ 26 %	+ 16 %
Pacific	Central Cities in Los Angeles PMSA	Renter	Black	+ 20 %	+ 10 %
Pacific	Non-Central Cities	Renter/ Owner	Black	+ 31 %	+ 6 %
Mid Atlantic	Non-Central Cities in Metro Areas w/ Central City > 250K	Renter/ Owner	Hispanic (except black)	+ 2 %	+ 16 %
Mid Atlantic	All Central Cities	Renter/ Owner	Hispanic (except black)	+ 14 %	+ 2 %
West South Central	Central Cities in Houston, Dallas, & Fort Worth PMSA's	Renter/ Owner	Hispanic (except black)	+ 8 %	+ 19 %
South Atlantic	All Non-Metro Areas & All Non-Central Cities	Renter/ Owner	Hispanic (except black)	+ 9 %	+ 22 %
West South Central	Central Cities in Metro Areas w/ Central City > 250K	Renter	White, Native Am., & Asian/Pacific except Hisp.	- 5 %	+ 11 %
East North Central	Central Cities in Metro Areas w/ Central City > 250K	Renter	White, Native Am., & Asian/Pacific except Hisp.	+ 21 %	+ 4 %
East North Central	Central Cities in Detroit and Chicago PMSA's	Renter	White, Native Am., & Asian/Pacific except Hisp.	- 4 %	+ 14 %
West South Central	Central Cities in Houston, Dallas, & Fort Worth PMSA's	Renter	White, Native Am., & Asian/Pacific except Hisp.	+ 7 %	+ 21 %
South Atlantic	Central Cities in Metro Areas w/o Central City > 250K	Renter/ Owner	White, Native Am., & Asian/Pacific except Hisp.	+ 10 %	- 1 %
South Atlantic	Non-Metro Areas Except Places > 10K	Renter/ Owner	White, Native Am., & Asian/Pacific except Hisp.	+ 3 %	+ 16 %

[14] U.S. Department of Commerce, Bureau of the Census. Unpublished file dated 6/14/91 containing adjustment factors derived from the 1990 Post-Enumeration Survey, prior to application of a statistical smoothing procedure. These adjustment factors reflect the amount of apparent net undercount actually measured in the PES sample for the indicated geographic areas and demographic groups.

Thus, the adjustment factors in Figure 3 suggest a high level of measurement error [15] rather than the high level of precision required for an adequate estimate of undercount.

Would the Adjustments Increase or Decrease Accuracy? The PES findings in Figure 3 provide a good basis for testing whether we can trust a coverage survey when it tells us that some population groups have higher undercounts than others. We have seen that these apparent undercounts seem to be implausible, but that by itself does not prove that they did not happen. If we can confirm that these differential undercounts did take place, then the credibility of coverage surveys as a tool for measuring undercount will be greatly increased. On the other hand, if it can be demonstrated that they did not take place, then the credibility of coverage surveys will be lost: If a coverage survey can indicate large undercount differentials where they do not exist, then it is obviously not a very reliable tool for measuring undercount.

Fortunately, because the ratio of male to female children is one of the most stable of all demographic statistics, these adjustment factors can be tested quite definitively. For each of the nation's nine regions, 51% of the young children enumerated in the 1990 Census were boys and 49% were girls. Likewise, for each of the major race categories, 51% of the young children enumerated were boys and 49% were girls. Among the nation's 284 metropolitan areas and consolidated metropolitan areas, the percent of young children who were boys varied very little, ranging from a low of 50.3% in Pine Bluff, Arkansas, to a high of 52.1% in Topeka, Kansas. Therefore, if the large differential undercounts indicated in Figure 3 really did take place, they should be very obvious: Boys should represent less than 51% of the total for areas with a large undercount of boys, but they should represent more than 51% of the total for areas with a large undercount of girls. Furthermore, if the undercounts indicated by the coverage survey really did take place, we should expect each area to move closer to the norm after it is "corrected" for Census undercount.

In fact, however, we find just the opposite. Figure 4 shows that the percentage of children under age 10 who are boys is about the same not only in each region, each race, and each metropolitan area, but also in the areas for which the coverage survey

[15] There are several types of measurement error. Although the point being made here is that the large amoun t of error in the adjustments is consistent with the thesis that large amounts of *non-sampling* error are inevitable, it should be noted that *sampling* error is also a very serious problem for the undercount adjustments. Actually, there is more than enough error to go around: these adjustments can reflect a very large amount of sampling error as well as a very large amount of non-sampling error. For purposes of data quality, both types of error are very problematic.

Figure 4

Before Adjustment for Undercount

Percent of Children Who Are Boys Offers No Surprises[16]

U.S. Total	51%
Race Categories:	
White	51%
Black	51%
Native American	51%
Asian/Pacific	51%
Other Race	51%
Hispanic	51%
Regions:	
New England	51%
Middle Atlantic	51%
East North Central	51%
West North Central	51%
South Atlantic	51%
East South Central	51%
West South Central	51%
Mountain	51%
Pacific	51%
Extremes Out of 284 Metro Areas:	
Lowest: Pine Bluff, Arkansas	50.3%
Highest: Topeka, Kansas	52.1%

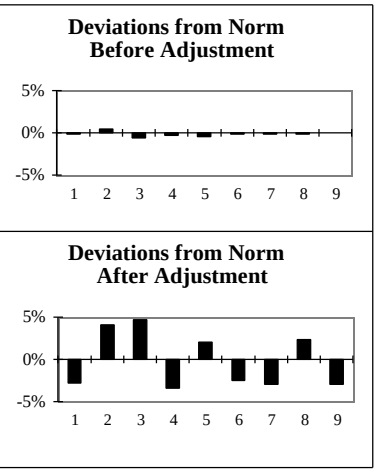
Selected Areas for Which the PES Indicated a Large Differential Undercount Between Boys and Girls:	
1. Asians/Pacific Islanders in non-central cities of the Pacific Region	51%
2. Asians/Pacific Islanders in central cities of the New York City PMSA	52%
3. Blacks in non-central cities of the Pacific Region	51%
4. Hispanics in non-central cities of large MSA's in the Mid-Atlantic Region	51%
5. Hispanics in central cities of the Mid-Atlantic Region	51%
6. Hispanics in central cities of the Houston, Dallas, and Fort Worth PMSA's	51%
7. Hispanics in non-central cities or non-metropolitan areas of the South Atlantic Region	51%
8. Non-Hispanic Whites, Native Americans, and Asians/ Pacific Islanders in central cities of small MSA's in the South Atlantic region	51%
9. Non-Hispanic Whites, Native Americans, and Asians/ Pacific Islanders in non-metropolitan areas of the South Atlantic region (excluding places with over 10,000 persons)	51%

Figure 5

After Adjustment for Undercount

Dramatic Variations in Percent of Children Who are Boys[17]

1. Asians/Pacific Islanders in non-central cities of the Pacific Region	49%
2. Asians/Pacific Islanders in central cities of the New York City PMSA	55%
3. Blacks in non-central cities of the Pacific Region	56%
4. Hispanics in non-central cities of large MSA's in the Mid-Atlantic Region	48%
5. Hispanics in central cities of the Mid-Atlantic Region	53%
6. Hispanics in central cities of the Houston, Dallas, and Fort Worth PMSA's	49%
7. Hispanics in non-central cities or non-metropolitan areas of the South Atlantic Region	48%
8. Non-Hispanic Whites, Native Americans, and Asians/ Pacific Islanders in central cities of small MSA's in the South Atlantic region	54%
9. Non-Hispanic Whites, Native Americans, and Asians/ Pacific Islanders in non-metropolitan areas of the South Atlantic region (excluding places with over 10,000 persons)	48%



[16] The percent of children who are boys was calculated based on the 1990 Census of Population and Housing, U.S. Department of Commerce, Bureau of the Census, Summary Tape File 1-C. Because Census counts by age, race, sex, and tenure have not been published, this table does not include the nine pairs of adjustments in Figure 3 which apply only to renters or only to homeowners. Although the race distinctions which are made in Summary Tape File 1-C do not correspond precisely to the race distinctions upon which the undercount adjustments were calculated, these discrepancies involve a very small number of people and they do not significantly affect the present analysis. Black Hispanics are counted as Hispanic in STF 1-C, but they should not be included with other Hispanics for purposes of applying undercount adjustments. Likewise, Asians/Pacific Islanders of Hispanic origin are counted as Asians/Pacific Islanders in STF 1-C, but they should not be included with that group for purposes of applying undercount adjustments.

[17] The data in Figure 5 were calculated after applying the adjustment factors from Figure 3 to Census counts from Summary Tape File 1-C

found large undercount differentials between boys and girls. It is only after applying these adjustments derived from the coverage survey that serious anomalies are found. As shown by Figure 5, the percentage of children who are boys deviates dramatically from the norm after adjustment. Even though Pine Bluff and Topeka are "outliers" among the nation's metropolitan areas, the adjusted Census counts are *two to six times* as far from the norm as Pine Bluff and Topeka. Thus, these "undercounts" measured in the PES sample do not correspond at all to actual undercounts in the areas which the sample represents. The Census is not really broken until after it is fixed.

The point being made here is not merely that the 1990 coverage survey produced faulty undercount measurements for young boys and girls. The problem is much broader than that, since the difficulties discussed in this paper apply just as much to other age groups as to children, and just as much to other demographic characteristics as to the sex ratio. The foregoing analysis focuses on the sex ratio of children merely because sex ratios provide a convenient and definitive basis for demonstrating the implausibility of the undercount measurements below the age where school attendance, military service, and employment patterns cause different communities to have a different mix of males and females. The focus on the sex ratio of young children should not by any means imply that undercount measurements are worse for this age group or that they would affect sex ratios more than the other population and housing characteristics that are measured by the Census. In the absence of any known problem that would scramble the undercount measurements for boys and girls without affecting the figures for other age groups and other demographic characteristics, we have to suspect that the measurements are faulty in other respects as well. The point being made is therefore nothing less than this: Because the large undercount differentials shown in Figure 3 are clearly spurious, we cannot trust a coverage survey to tell us which segments of the population have higher undercounts than others.

Does It Make a Difference? It may take a few moments to comprehend the impact that adjustment factors like those displayed in Figure 3 would have if they were applied to the Census.[18] To those of us who have become accustomed to Census

[18] The adjustment factors in Figure 3 reflect the amount of apparent net undercount actually measured in the PES sample for the indicated geographic areas and demographic groups. It should be noted that these factors were subsequently subjected to a statistical "smoothing" procedure to produce new factors that followed a more consistent pattern by age, race, and sex. It was these "smoothed" factors that were actually proposed in 1991 for use in adjusting the 1990 Census. Further modifications proposed in 1992 for use in adjusting the population base for population estimates would have combined males and females under age 17. The resulting "collapsed" adjustment factors represent the Census Bureau's latest official estimate of undercount in the 1990 Census. The "smoothed" adjustment factors would be appropriate for use in estimating the practical impact of adjusting the 1990 Census data for undercount. The "unsmoothed" adjustment factors are pertinent for the current analysis, since they reflect the amount of apparent undercount actually identified by the PES. The unsmoothed factors are

data that generally make sense at the local level, it is mind-boggling to consider the prospect of largely arbitrary adjustments—and sometimes arbitrarily large ones—applied to every number in the Census. In an effort to address a relatively small inaccuracy at the national level, we would utterly destroy the reliability of Census data at the state and local level.

Perhaps most alarming is the impact on comparisons over time. If coverage surveys can indicate large differential undercounts between boys and girls even where no differences exist, they can also indicate large differential undercounts between one Census and the next where no differences exist. To illustrate the potential implications of this problem, let us consider what would happen if there turns out to be no real difference in certain undercount rates for Census 2000 and Census 2010, but the coverage surveys indicate the same spurious differences between these two points in time that the 1990 PES found between boys and girls. Under these assumptions, the numbers in Figure 3 could all remain the same, [19] but they would represent spurious undercount differentials between Census 2000 and Census 2010 instead of spurious undercount differentials between boys and girls in 1990. This would generate many interesting demographic "findings":

- The counts of Asians/Pacific Islanders in non-central cities of the Pacific region would be inflated by 5% in 2000 but by 17% in 2010. (See line 1 of Figure 3.) The adjusted Censuses would therefore suggest far greater growth in the number of Asians than actually occurred. What effect would this have on attitudes toward Asian immigrants in these communities?

- The count of black homeowners in central cities of the Los Angeles PMSA would be inflated by 28% in 2000 but by only 8% in 2010. Similarly, the count of black renters would be inflated by 20% in 2000 and by 10% in 2010. (See lines 4 and 7 of Figure 3.) The adjusted Census data would therefore show a large exodus of the black population and a substantial drop in black home ownership for Los Angeles relative to the actual trend. What impact would this have on race relations? What would be the impact on government housing programs and anti-discrimination programs?

---

also relevant in the context of Census 2000, since the Census Bureau does not plan to use a statistical smoothing process in the next Census.

[19] Our assumption that "the undercount adjustments indicate the same spurious *differences* between these two points in time that the 1990 PES found between boys and girls" does not require the adjustments themselves to be the same as the 1990 adjustments for boys and girls, but merely for the differences to be the same. The numbers "could" remain the same, but they would not necessarily have to. For simplicity and clarity of presentation, the illustrations are based the special case in which the adjustments are the same.

- The count of black homeowners in central cities of the New York City PMSA, on the other hand, would be inflated by 0% in 2000 and by 23% in 2010. (See line 5 of Figure 3.) This area would therefore seem to have a dramatic rise in black home ownership relative to the actual trend. Of course, home ownership would not by any means be the only variable affected by these faulty adjustment factors: Poverty, marital status, and every other characteristic that is correlated with race and with home ownership would also be affected. Social scientists could spend the decade trying to explain why the economic status of blacks seemed to rise so rapidly in New York city while it seemed to decline in Los Angeles. What would be the impact on the credibility of the Census when they discovered the answer?

- The counts of White, Native American, and Asian/Pacific renters in Detroit and Chicago would be *decreased* by 5% in 2000, but they would be inflated by 11% in 2010. Thus, there would seem to be a dramatic increase in renters and a shift away from home ownership in these cities relative to the actual trend. (See line 15 of Figure 3.) In contrast, other central cities in these same metropolitan areas would have their counts for these demographic categories inflated by 21% in 2000 and by only 4% in 2010. (See line 14 of Figure 3.) The faulty adjustment factors would therefore make it appear that huge numbers of white renters had moved from Detroit and Chicago to other nearby central cities before 2000, but that they moved back in the next decade.

Of course, these illustrations are only hypothetical. Perhaps Los Angeles will have reasonable undercount adjustments for black homeowners in 2000 and 2010. Maybe its adjusted Census data will show a spurious decline in its elderly population instead, and maybe it will be New York that shows a spurious decline in black home ownership. We won't know before it happens. Even worse, we won't know even *after* it happens. When adjusted Census data suggest a dramatic change in population trends, we will not know how much of the change represents actual demographic shifts and how much represents spurious differences in undercount adjustments. Are we ready to discover dramatic new (and totally false) trends in disease prevalence, mortality rates, school enrollment, income distribution, housing patterns, marital status, welfare dependency, gender differences, and all of the other issues that are studied on the basis of Census data? We expect a Census to increase our knowledge about population trends, but an adjustment methodology which can indicate large differentials where differentials do not exist would increase our ignorance instead.

Conclusion. We cannot escape the conclusion that the method proposed for correcting Census undercount has some rather serious shortcomings. The impact on the validity of the 1990 Census would have been devastating, and we can expect the impact on Census 2000 to be similar: The problems are not due to minor flaws in methodology or implementation, but rather to the impossibility of measuring undercount through the proposed coverage survey. Unless we can convince people who don't want to be counted to answer our surveys, and unless we can replicate and match the valid Census results with near-perfect accuracy, any undercount estimates that are developed in this manner will be dominated by measurement error. Instead of describing variations in the amount of undercount from one area to another, they will largely describe variations in the amount of error in replicating the Census and in matching individuals identified by the survey with individuals identified by the Census. Once the impossibility of the task is recognized, one can only be impressed by how close the Census Bureau seemed to come to succeeding in 1990. However, one must also be impressed by how close we are to destroying the credibility and the value of the Census.

Quantifying Measurement Error and Bias in the 1990 Undercount Estimates

Kenneth Darga, Senior Demographer
Office of the State Demographer
Michigan Information Center
Michigan Department of Management and Budget

April 29, 1998

Please address comments and suggestions regarding this analysis to:
DargaK@state.mi.us

The opening pages of the preceding paper[1] set up a paradox: Since the number of people who want to avoid being identified by the government is more than sufficient to account for the level of undercount identified through demographic analysis, and since many of these people can be counted upon to avoid the coverage survey as well as the Census, how is it that the 1990 coverage survey suggests about the right level of total undercount at the national level?

The solution I have proposed is that this "correlation bias"—i.e. missing many of the same people in both the coverage survey and the Census—is offset by counting some people as missed by the Census when they really were included. I have suggested that, rather than just reflecting undercount, the undercount factors derived from the coverage survey reflect a variety of methodological difficulties involving imperfect replication of the census, survey matching, unreliable interviews, geocoding problems, and the like.

The preceding paper demonstrates that this is a *plausible* solution to the paradox and that it is *consistent* with both the plausible undercount estimates at the national level and the implausible estimates for individual poststrata. It shows that, although an extremely high level of accuracy is required for an adequate measure of undercount, the obstacles to an accurate coverage survey are immense. It points out many specific types of error that are difficult or impossible to avoid, and it shows that the proposed undercount adjustments for 1990 were suggestive of high levels of error.

Even these limited accomplishments of the paper are significant: Proponents of the proposed undercount adjustment are left with the task of explaining how the 1990 coverage survey could indicate very large and demonstrably spurious differential undercounts for young children. In addition, they must explain how we can rely upon the 5%, 10%, and 20% differential undercounts identified between other poststrata when the 5%, 10%, and 20% differential undercounts identified between young boys and girls are known to be spurious. They must make a believable argument that the coverage survey somehow really did count critical groups of people who were missed by the 1990 census, i.e. homeless people and the illegal immigrants, drug dealers, fugitives, and others who don't want the

[1] Kenneth J. Darga, "Straining Out Gnats and Swallowing Camels: The Perils of Adjusting for Census Undercount," Office of the State Demographer, Michigan Information Center, Michigan Department of Management and Budget, 1998.

government to know where they are. They must demonstrate either that they achieved extremely low error rates in the face of seemingly insurmountable obstacles, or else that—notwithstanding the demonstrated inaccuracies of the undercount measurements for some individual poststrata—they have enough luck and skill to ensure that large errors will offset each other very precisely. Merely a general tendency for errors to offset one another is not enough: An extremely high level of accuracy is required to measure a phenomenon as small and elusive as census undercount at the sub-national level. Each of these issues is critical to the success of the effort to measure undercount. The credibility of the proposed method cannot be restored unless its proponents are successful on *all* of these points.

A major limitation of the preceding paper is that, although it suggests what sorts of errors are difficult or impossible to avoid, it stops short of showing that those errors actually occurred or how serious they were. To fill this gap in the analysis, this paper relies upon evaluation studies by the Census Bureau and the work of other analysts. That work confirms that the errors are very large indeed, and that they did not offset each other precisely in the analysis of the 1990 coverage survey.

The Census Bureau has extensively evaluated the process and results of the 1990 coverage survey, which is commonly referred to as the "Post-Enumeration Survey" or "PES." Its findings are written up in 22 unpublished reports, eight of which are referenced in this paper. These reports, which are known as the "P-project reports," were issued in July 1991 under the main title "1990 Post-Enumeration Survey Evaluation Project." These reports are referred to in this paper by their number within the series, e.g. "P-4" or "P-16." Most of the references to these reports and many of the other quantitative observations which appear below are based upon the work of Dr. Leo Breiman, an emeritus professor of statistics at the University of California, Berkeley (Breiman, 1994).

Six major sources of error are quantified below: matching error, fabrication of interviews, ambiguity or mis-reporting of usual residence, geocoding errors, unreliable interviews, and the number of unresolved cases. It will be seen that the level of error and uncertainty contributed by each of these factors is very substantial relative to the magnitude of net undercount. Thus, each of these error sources by itself is sufficient to demonstrate that the sort of coverage survey used by the Census Bureau is not capable of accurately measuring Census undercount. It will then be shown that the various identified sources of error actually did increase the 1990 undercount estimate enough to explain the paradox.

1. Matching Error

A critical step in measuring undercount through a coverage survey is to match people counted in the coverage survey with people counted in the Census. Most people are counted by both surveys, but problems such as misspellings, misreporting of age, language barriers, aliases, missing data, errors in recording the address, changes in household composition, and a host of other difficulties can make it difficult to match up the records. Any failure to match the records can lead to an overestimate of undercount: The person's record in the Post-Enumeration Survey—sometimes referred to as the "P-Sample"—can be mistakenly counted as having been missed by the Census. Yet their Census response—the Census enumerations from the same geographic areas are sometimes referred to as the "E-sample"—cannot be classified as erroneous unless strict criteria are met.[2] (After all, it is a valid record.) Thus, when records fail to match, it is possible for people to be counted twice. The many barriers to matching the coverage survey results with the Census are described in a sidebar of the preceding paper, and their seriousness is confirmed by the results of the Census Bureau's evaluation studies.

As explained in the P-8 report, a computer-matching process was able to resolve about 75% of the P-sample records, and the remaining records went to two independent teams of trained matchers. Although these teams used the same definitions and guidelines, they had a surprisingly high rate of disagreement regarding which people counted by the PES had been counted by the Census. Of people classified as "matched" by the first team, 5.7% were classified as "not matched" and 4.5% were classified as "unresolved" by the second team. Of those classified as "not matched" by the first team, 4.8% were classified as "matched" and 1.3% were classified as "unresolved" by the second team. Of those classified as "unresolved" by the first team, 22.7% were classified as "matched" and 8.0% were classified as "unmatched" by the second team. (Ringwelski, 1991). Although the matching process must achieve near-perfection in order to accurately measure the 1% or 2% of the population that is missed by the Census, it is obviously a very difficult task, and even teams using the same guidelines can differ widely in their judgments.

---

[2] For example, Howard Hogan, then director of the Undercount Research Staff of the Census Bureau, wrote: "Proving that someone does not exist is not easy. . . . The rules require the interviewer to find at least three knowledgeable respondents in an effort to determine whether an enumeration was fictitious." (Hogan, 1991a). This would be difficult to do in a case where an unmatched person really existed.

This high level of disagreement has several serious implications:

- First, it indicates that the number of "difficult" cases for which match status is not obvious is very large, greatly exceeding the estimated level of net undercount. This demonstrates the impossibility of measuring undercount accurately through a coverage survey even apart from any other considerations.

- Second, since trained teams differ substantially in their judgments, it follows that some of the judgments reached by the final team of matchers are likely to be wrong: Some of the people counted by the Census will be identified as missed, some of the people missed by the Census will be identified as counted, some of the people counted correctly by the census will be identified as counted in error, and some of the people counted in error will be identified as counted correctly. If the number of difficult cases were small, we could hope that the errors would come close to cancelling each other out. However, given the high level of disagreement between the matching teams, any of these types of error could potentially exceed the actual level of undercount: "close" is therefore not enough.

- Third, since high levels of subjectivity and art are obviously involved in the matching process, it is subject to additional sources of bias. Will the match rate be different if the cases are examined in the first week of matching or in the final week? Will the match rate be different depending on which regional office examines them? If a difficult case falls into a category that is expected to have a high undercount rate, will that decrease its likelihood of being classified as matched? If a similar case falls into a category that is expected to have a low undercount rate, will that increase its likelihood of being classified as matched? Such issues can have a significant impact on the differential undercount rates of individual poststrata and of different geographic regions. If matching were an objective process whose results could be fully determined by the Census Bureau's matching rules, these questions would be insignificant. However, because the process is obviously a somewhat subjective one, these questions become very important. In fact, since the number of difficult cases is quite large and the level of disagreement between teams exceeds the total level of undercount, these questions must be considered critical.

- A fourth implication of the high level of disagreement between different match teams is that the results for a given set of records are likely to be different each time the match is performed. Clear evidence of this is provided by the results of rematching selected blocks which initially had large numbers of non-matches

and erroneously enumerated persons: Rematching only 104 out of the 5,290 block clusters resulted in a decrease of 250,000 (about 5%) in the estimated net national undercount. (Hogan, 1993).

2. Fabrication of Interviews

The problem of fabricated data is another example of a data collection problem whose magnitude is very substantial relative to the magnitude of Census undercount. Many large surveys conducted by the Census Bureau appear to have a significant number of records that are fabricated by the interviewer. Previous research has shown that, overall, between 2% and 5% of the interviewers are dishonest in their data collection and that between 0.5% and 1.5% of the interviews themselves are fabricated (Stokes and Jones, 1989). One-time surveys such as the Census and the PES are particularly vulnerable to this problem, since temporary employees are found to be more likely to fabricate records than permanent employees. Workers who are detected fabricating data sometimes do so on a large scale. Biemer and Stokes (1989) found that, on average, inexperienced interviewers who were detected fabricating data did so for 30% of the units in their assignment; for more experienced interviewers, the rate was 19%.

While the prospect that perhaps 0.5% or 1.5% of the Census and PES interviews are fabricated may not sound extremely serious at first, it must be remembered that we are trying to measure a net undercount of only about 1% or 2% of the population. Thus, instead of saying that 0.5% and 1.5% are small relative to 100%, it is more pertinent to say that they are very substantial relative to 1% or 2%. (Of course, it should be noted that undercount rates are higher than 1% or 2% for some demographic groups and some types of area. However, that does not greatly affect this comparison, since fabrication rates also tend to be highest in the areas that are most difficult to enumerate. See Tremblay, 1991, and West, 1991c).

Both fabrication in the Census and fabrication in the PES have very serious implications for estimating undercount. When a block cluster with interviews that were fabricated by a Census enumerator is included in the PES, it will raise the rates of undercount and erroneous enumeration for the poststrata represented within it. Since, as already noted, it is difficult to prove that people do not exist, the increase in the apparent rate of erroneous enumeration may not be as great as the increase in the apparent undercount rate. This would lead to an overestimate of net undercount for these poststrata. Fabrication within the PES is even more

problematic. When people counted by the PES are matched against Census questionnaires, any fabricated PES records can look like people who were missed by the Census. However, when the corresponding Census records are tested for validity, they are likely to be classified as valid: It is particularly difficult to prove that someone does not exist if they really do exist. Thus, fabrication once again can lead to an overestimate of net undercount. Fabricated PES records would be particularly difficult to detect in cases where the housing unit was vacant during the Census or during PES follow-up.

The actual amount of fabrication in the PES is difficult to determine. The P-5a report, which is based on data which were not specifically designed to detect fabrication, identified only 0.03% of the cases in the P-sample evaluation follow-up data as fabrications (West, 1991b). These cases were estimated in the P-16 report to have inflated the national undercount estimate by 50,000 persons, or about 1% of the total net undercount (Mulry, 1991). The P-5 report, on the other hand, used quality control data collected during the PES to identify 0.26% of the PES household interviews and 0.06% of the remaining cases on a national level as fabrications (Tremblay, 1991). Although this is a much lower rate of fabrication than would be expected based on the studies cited above, it is nevertheless about eight times the proportion of cases identified as fabrications in the P-5a report, suggesting that perhaps fabrications represent about 8% of the total net undercount. Yet another Census Bureau report on this issue, the P-6 report, was designed to gain knowledge about fabrication that may have been undetected in the quality control operation. This report found that only 39% of the interviewers whose match rates were suggestive of high levels of fabrication had been identified in the quality control operation. (West, 1991c). This suggests that the level of fabrication in the PES may have been close to the level that has been found in other similar surveys, making it a very significant problem indeed.

The P-6 report also found that fabrication rates seemed to vary substantially from one region to another. Interviewers who appeared to have high levels of fabrication accounted for 2% to 5% of the interviews in most regions, but they accounted for 7.7% of the interviews in the Atlanta regional office and 8.8% of the interviews in the Denver regional office (West, 1991c.). Regional variation in the amount of fabrication is not surprising, since important factors which are likely to influence the fabrication rate vary by region. For example, while PES interviews to identify undercount were being conducted at the end of June and into July of 1990, most of the northeast and midwest had very pleasant weather. Much of the

south and west, on the other hand, had long periods with temperatures near or above 100 degrees. Denver, for example, had eleven consecutive days at the end of June and the beginning of July with temperatures of 95 degrees or higher, including five days with temperatures in the 100's. Atlanta had seventeen consecutive days with temperatures of 89 degrees or higher, followed by several days of rain. Thus, it is not surprising that fabrication seems to have been a more serious problem in these areas. Moreover, since fabrication also varies substantially by neighborhood, with interviewers being more likely to fabricate records in neighborhoods they perceive as dangerous than in safer neighborhoods, it also varies by race and by owner/renter status. It therefore appears that fabrication can account for a substantial portion of the undercount differentials identified between regions, between types of city, and between population groups.

3. Ambiguity or Misreporting of Usual Residence

The question of where someone lives is often not as straightforward as it may seem. The Census uses the concept of "usual" address: If you are staying somewhere temporarily and usually live somewhere else, you are instructed to report your "usual" address instead of your address on April 1. For many people, this instruction is ambiguous and subject to varying interpretation. "Snowbirds" who migrate between the north and south can give the address where they spend the largest part of the year, the address where they spend the largest part of their life, the address where they are registered to vote, the address where they feel most at home, or the address where they happen to be on April 1. They might give one answer when they fill out their Census form in April and a different answer when they are interviewed for the coverage survey in July. Other people who move to or from temporary quarters at about the time of the Census can also claim a "usual" address different from the place where they were located on Census day. For example, college students who are packing up to move out of a dormitory room that they will never see again may use their "home" address instead of the college address that the Census Bureau would prefer. In comparison with an estimated national undercount of only 1% or 2% of the population, these components of the population with an indistinct "usual" place of residence represent a very significant component of the population.

Thus, the task of determining the "appropriate" address for each Census respondent amounts to replacing the traditional concept of "usual" address, which is defined largely by the respondent, with a set of assignment rules developed by the designers of the coverage survey. This can involve the reassignment of large

numbers of people, and it can potentially have a larger impact on regional population distribution than Census undercount itself.

Given the large number of people with an indistinct "usual" place of residence, it not surprising that the Census Bureau's Evaluation Follow-Up Study found many P-sample respondents who were classified as non-movers for purposes of calculating the undercount adjustments, but were identified by new information as having moved in after census day. Weighted to a national level, they represented 274,000 persons,[3] or about 5% of the estimated national net undercount. (Of course, the impact on the individual poststrata that were most affected would have been greater.) It should be noted that these figures do not reflect the full magnitude of the problem of indistinct "usual" place of residence: they reflect only those cases—presumably a small minority—for which the PES was judged to have classified movers incorrectly.

Finally, it should be noted that different cities and different neighborhoods can vary greatly in their proportion of people with an indistinct "usual" place of residence. If the sample drawn for particular poststratum happens to include some block clusters in a college town or in a retirement community, then its adjustment factor will be very strongly affected by this problem. The adjustment for a class of cities in an entire region can thus be determined largely by whether or not the sample includes a few "outlier" blocks.

4. Geocoding Errors

Another task which proves to be very difficult is coding addresses to the proper Census Block. Coding a record to the wrong Census block is a very serious problem for an undertaking that depends upon matching records between two surveys. If a Census record that belongs in a sample block has been mistakenly coded to a different block, it may not be found. The corresponding PES record would therefore be erroneously classified as missed by the Census. On the other hand, if an otherwise valid Census record has been mistakenly coded to the sample block, it may be counted as an erroneous enumeration when it fails to match with a

[3] The P-4 report (West, 1991a) and P-16 report (Mulry, 1991) indicated that "census day address error" increased the undercount estimate by 811,000 persons. However, the Census Bureau subsequently indicated that this figure included other errors found by the P-sample re-interview as well (Breiman, 1994, p.475). The conclusion that 274,000 persons were found to have been added to the undercount estimate through incorrect assignment of Census-day address by the PES is based on subtracting these other errors, which represent 537,000 persons labeled "P-sample re-interview" in Dr. Breiman's paper, from the 811,000 persons initially identified as "census day address error" in the Census Bureau reports. (See Breiman, 1994, pp.467, 471, and 475.)

PES record and when residents of the block indicate that no such person lives there. To reduce the magnitude of these problems, both PES records in the P-sample and Census records in the E-sample were checked against one or two rings of surrounding blocks. According to the P-11 report, 4.08% of the P-sample was matched to the Census through geocoding to the surrounding blocks, but only 2.29% of the E-sample was classified as correctly enumerated as a result of matching with PES records in surrounding blocks. If matching to surrounding blocks had not been done, this difference would have been equivalent to an approximate excess of 4,296,000 in the P-sample population (Parmer, 1991, Attachment).

This difference highlights the sensitivity of the PES analysis to variations in methodology and procedure. As pointed out by Dr. Leo Breiman: "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 . . . national estimated undercount to over 4%. On the other hand, using a larger search area might well have produced a much lower undercount estimate." (Breiman, 1994, p.468.) Since 38% of the households that were matched outside their proper block in the 1986 PES rehearsal were matched more than five blocks away (Wolter, 1987), an expanded search area might have had a very significant effect on the measure of undercount.

The sensitivity of the PES analysis to small variations in methodology and procedure is also illustrated by another geocoding problem encountered by the PES. It was found that two particular block clusters initially increased the undercount estimate by nearly one million people due to faulty census geocoding. Most of the people in those blocks had been counted by the Census, but many of them were identified as uncounted because they had been erroneously coded as living in different blocks. It is somewhat disconcerting that only two block clusters out of a total of 5,290 included in the PES can erroneously contribute nearly one million people to the undercount estimate, especially since the total estimated net undercount is only about five million. Of course, in this case the problem was obvious enough to be identified: the influence of these block clusters was downweighted so that they contributed "only" 150,000 to the estimated undercount. (Hogan, 1991b). One has to wonder, however, how many similar problems may have gone undetected and uncorrected.

5. Unreliable Interviews

Another problem which the PES must contend with is unreliable interviews. Interviews can be unreliable for many reasons, including interviewer errors,

language barriers, lack of information on the part of respondents (some of whom are children and some of whom are neighbors, landlords, or other non-members of the household), and lack of cooperation on the part of respondents (some of whom are criminals, illegal immigrants, psychotics, or practical jokers). The serious implications of this problem for measurement of undercount through a coverage survey are demonstrated in the P-9a report. The Evaluation Follow-Up project conducted new interviews for a sample of PES E-sample records. The new interview information was given to matching teams with instructions to change match status only if new, relevant, and reliable information was present in the new interview. The result was that 13% of the records changed match status. In fact, a majority of these changes (7% of the records examined) involved changes from "erroneous enumeration" to "correct enumeration" or vice versa; the remainder (6% of the records examined) involved changes from one of these categories to "unresolved" or vice versa (West, 1991d; Ericksen *et. al.*, p.512). Although Ericksen et. al. stress the fact that the changes had a general tendency to cancel each other out and that they had fairly little effect on the net undercount estimates, the more pertinent implication for the present analysis is that a very substantial proportion of cases from the Post-Enumeration Survey had very uncertain match status. Whether these changes in match status are attributable to unreliable information in the initial interviews or merely to a tendency for match status to change each time a different team of matchers examines a difficult case, the fact remains that we are trying to measure a subtle phenomenon with a very crude instrument. Based on the findings in the P-9a report, weighted to reflect the national population, over 2 million persons would have changed from "correctly enumerated" to other classifications, and over 1.6 million persons would have changed from "erroneously enumerated" to other classifications (West, 1991d). In the context of a net national undercount of only about 5 million people, the magnitude of these reclassifications suggests very serious problems resulting from unreliable interview data.

6. Unresolvable Cases

After all of the followup, review, and rematching involved in the 1990 PES, there were still 5,359 E-sample cases and 7,156 P-sample cases which remained unresolved and had to be imputed. This represents approximately 1.6% of the total combined P-sample and E-sample cases. On the one hand, the fact that the number was not larger is a testimony to the persistence and ingenuity of the PES staff. On the other hand, it must be noted that the percentage of unresolved cases was very close to the total percentage of the population that is believed to be undercounted.

Thus, unresolved cases are not a small problem, but rather a problem that can have a critical impact on the undercount estimate. As Dr. Breiman notes, the undercount estimate would nearly double if all of the unresolved P-sample cases were assumed to be unmatched and all of the E-sample cases were assumed to be correctly enumerated, but the opposite assumptions would suggest a census *overcount* of one million persons (Parmer, 1991; Breiman, 1994, p.468).

The match status of the unresolved cases was imputed through a complex regression model that involved estimating coefficients for dozens of variables (Belin, et.al., 1993). However, regardless of the complexity of the methodology or the carefulness of its assumptions, it must be recognized that the cases we are talking about here are all ones that could not be classified as matches or non-matches even after careful and repeated review of all of the information available about them. Very little is known about what proportion of unresolvable survey responses really do match with one another. An imputation process may be able to produce a "reasonable" allocation of records to matched and unmatched status, but it cannot classify them definitively. A "reasonable" allocation would be sufficient if the proportion of unresolved cases were very small relative to the rate of undercount, but it is not sufficient when the proportion of unresolved cases is nearly as great as the net rate of undercount. The large number of unresolvable cases is by itself a fatal flaw in the undercount analysis.

Impact of Identified Sources of Error on the Undercount Estimate

We have seen that the undercount measurements are subject to several serious sources of error. In order to determine whether these errors can serve as a solution to the paradox identified at the beginning of this paper, it is necessary to see whether their combined effect would elevate the undercount estimates enough to offset the tendency for the coverage survey to miss many of the same people that are missed in the Census.

Several attempts have been made to quantify the net effect of identified measurement errors on the 1990 estimates of undercount. The analysis in the Census Bureau's P-16 report indicates that corrections for measurement errors in the 1990 PES would have decreased the undercount estimate from 2.1% to 1.4% (Mulry, 1991). A later analysis by the same author incorporated additional corrections related to a major computer processing error discovered by the

Figure 1
Impact of Identified Sources of Error on the 1990 Undercount Adjustments

Error Source	Impact on Undercount Estimate[4] (i.e. number of persons erroneously added to undercount)
P-sample rematching	553,000
Census-day address errors	274,000
Fabrications	50,000
E-sample rematching	624,000
E-sample re-interview	-473,000
P-sample re-interview	537,000
Ratio estimator bias	290,000
Computer coding error	1,018,000
Late-late Census data	183,000
New out-of-scopes in re-match	164,000
New out-of-scopes in re-interview	358,000
Re-interview of non-interviews	128,000
TOTAL	3,706,000
Estimate of identified net undercount prior to correction for identified errors:	5,275,000
Estimate of identified net undercount after correction for identified errors:	1,569,000

Note: The first seven of these error sources are considered in the P-16 report (Mulry, 1991), and the first nine error sources are considered in the subsequent Census Bureau report by the same author (Mulry, 1992).

[4] With the exception of the count of Census day address errors, these figures are taken from Table 15 of Breiman (1994). That table indicated 811,000 Census day address errors, based on the P-4 and P-16 reports. As explained in Footnote 3 above, that figure is corrected here to 274,000. This correction is also reflected in Dr. Breiman's finding that correction of identified errors would lower the undercount estimate to 0.6%. Excluding that correction, Dr. Breiman's adjusted undercount estimate was only 0.4%.

It should be noted that, like the original PES estimates of undercount, these estimates of PES error are subject to both sampling error and non-sampling error. Moreover, it is likely that they fail to identify all of the problems of the PES. Nevertheless, these estimates are more than adequate for the present purpose of demonstrating that the 1990 coverage survey involved a very large amount of measurement error and that its identified errors are sufficient to explain the paradox laid out at the beginning of this paper. However, they should not be interpreted as producing a definitive estimate of the amount of "true" undercount that was identified by the 1990 PES.

Census Bureau in late 1991, the rematching of records in some suspect blocks, and the inclusion of very late Census data that had not been available when the initial PES estimates were developed. This analysis suggested that corrections for identified measurement errors would have reduced the undercount estimate from 2.1% to 0.9% (Mulry, 1992). An analysis by Dr. Leo Breiman, which built upon the Census Bureau analyses cited above, incorporated additional sources of error to arrive at an adjusted undercount estimate of only 0.6% (Breiman, 1994, p.475). This does not mean that the "true undercount" was only 0.6%, but merely that this is the amount of apparent undercount identified by the 1990 coverage survey which remains after making rough adjustments for the errors that have been identified and documented. Dr. Breiman's estimates of the impact of each error source, based on data from the Census Bureau evaluations, are shown in Figure 1. Dr. Breiman concludes that about 70% of the net undercount adjustment that had been proposed for the 1990 Census count—3,706,000 out of 5,275,000 persons—actually reflects identified measurement errors rather than actual undercount.

Despite their differences, these three studies all point clearly to the same conclusion: There are enough measurement errors which inflate the undercount estimate to roughly offset the large number of people who appear to be missed by both surveys. This provides the solution for the paradox identified at the beginning of this paper.

Thus, it appears that the 1990 coverage survey missed a very substantial number of people who were missed by the Census, but that it also identified a large number of people as missed by the Census who actually had been counted. Moreover, there is a large amount of additional error—far greater in magnitude than the level of undercount—which is less visible at the broadest level of aggregation because the errors in one direction are offset by errors in the other direction. Thus, while the 1990 coverage survey suggests an overall level of undercount similar to that indicated by demographic analysis, it cannot be relied upon to shed light on patterns of undercount for different demographic components of the population or for different geographic areas. The differential undercounts indicated by the coverage survey largely reflect differences in the incidence and direction of survey matching errors and other methodological problems rather than differences in the incidence of Census undercount. As noted in the preceding paper, this does not reflect deficiencies in the skill and effort applied to the task by the Census Bureau, but rather it reflects the impossibility of adequately measuring undercount in this manner.

REFERENCES

Belin, T.R., G.J. Diffendal, S. Mack, D.B. Rubin, J.L. Schafer, and A.M. Zaslavsky (1993). Hierarchical Logistic Regression Models for Imputation of Unresolved Enumeration Status in Undercount Estimation. *Journal of the American Statistical Association,* 88:1149-1166.

Biemer, P.P., and Stokes, S.L. (1989). The Optimal Design of Quality Control Samples to Detect Interviewer Cheating. *Journal of Official Statistics,* 5, 23-39.

Breiman, Leo (1994). The 1991 Census Adjustment: Undercount or Bad Data. *Statistical Science,*9(4):458-537.

Hogan, Howard (1991a). The 1990 Post-Enumeration Survey: An Overview. Unpublished paper, U.S. Bureau of the Census.

Hogan, Howard (1991b). Downweighting Outlier Small Blocks. STSD 1990 Decennial Census Memorandum Series #V-109, addressed to John Thompson, Chief, Statistical Support Division, Bureau of the Census, Washington D.C., June 18, 1991.

Hogan, Howard (1993). The 1990 Post-Enumeration Survey: Operations and Results.*Journal of the American Statistical Association,* 88(423):1047-1060.

Mulry, Mary H. (1991). P-16 Report: Total Error in PES Estimates by Evaluation Post Strata. U.S. Bureau of the Census, 1990 Post-Enumeration Survey Evaluation Project, Series #R-6.

Mulry, Mary H. (1992). Total Error of Post Census Review Estimates of Population. Decennial Statistical Studies Division, U.S. Bureau of the Census, Washington D.C., July 7, 1992.

Parmer, Randall (1991). P-11 Report: Balancing Error Evaluation. U.S. Bureau of the Census, 1990 Post-Enumeration Survey Evaluation Project, Series #M-2.

Ringwelski, Michael (1991). P-8 Report: Matching Error—Estimates of Clerical Error from Quality Assurance Results. U.S. Bureau of the Census, 1990 Post-Enumeration Survey Evaluation Project, Series #I-2.

Stokes, S.L. and Jones P.M. (1989). Evaluation of the Interviewer Quality Control Procedure for the Post-Enumeration Survey, *American Statistical Association 1989 Proceedings of the Section of Survey Research Methods,* 696-698.

Tremblay, Antoinette (1991). P-5 Report: Analysis of PES P-Sample Fabrications from PES Quality Control Data. U.S. Bureau of the Census, 1990 Post-Enumeration Survey Evaluation Project, Series #E-4.

West, Kirsten K. (1991a). P-4 Report: Quality of Reported Census Day Address. U.S. Bureau of the Census, 1990 Post-Enumeration Survey Evaluation Project, Series #D-2.

West, Kirsten K. (1991b). P-5a Report: Analysis of P-Sample Fabrication from Evaluation Follow-Up Data. U.S. Bureau of the Census, 1990 Post-Enumeration Survey Evaluation Project, Series #F-1.

West, Kirsten K. (1991c). P-6 Report: Fabrication in the P-Sample: Interviewer Effect. U.S. Bureau of the Census, 1990 Post-Enumeration Survey Evaluation Project, Series #G-2.

West, Kirsten K. (1991d). P-9a Report: Accurate Measurement of Census Erroneous Enumerations. U.S. Bureau of the Census, 1990 Post-Enumeration Survey Evaluation Project, Series #K-2.

Wolter, Kirk M. (1987). Technical Feasibility of Adjustment. Memorandum to the Undercount Steering Committee, Bureau of the Census, Washington D.C.