Policy Forum

Accounting for America's Uncounted and Miscounted

KIRK M. WOLTER

THE DECENNIAL CENSUSES IN THE UNITED STATES MISS some people; others are counted more than once or in the wrong geographic location. The difference between the true, but unknown population count and an original census count is called the net undercount. In this article, I present evidence about the size of the net undercount, explain how it is measured, explain why it is an important problem, and demonstrate new statistical methodology that can ameliorate the problem.

Since 1940, the Census Bureau has prepared formal estimates of the percent net undercount (Table 1). These data reveal several important facts, including (i) censuses have counted most people; (ii) with the possible exception of 1990, each census has displayed improvement, in the sense of lower net undercount, than the previous census; (iii) men tend to be missed at a higher rate than women; (iv) blacks tend to be missed at a higher rate than nonblacks; and (v) notwithstanding the general improvement in coverage since 1940, the difference between the black and nonblack undercount rates has hovered in a constant range about 3.5 to 4.5%.

Additional data clearly show that the differential undercount phenomenon affects the highly mobile, young adults, the poor, Hispanics and other minority communities, and household residents who are not members of the nuclear family, among others. It varies locally; some localities suffer much higher rates than the nation as a whole.

If the rate of net undercount were nearly constant throughout the land, few people would be concerned. That people and localities are subject to differential undercount rates, however, is a profoundly important problem of national scope, directly tied to three basic uses of census data—apportionment, redistricting, and fund allocation.

The decennial census is mandated by the U.S. Constitution. Under the Constitution, the number of seats in the House of Representatives apportioned to each state is based on its population, as determined by the census. Under the constitutional principle of "one person, one vote." the U.S. Supreme Court has held that, within any state, congressional districts must "as nearly as is practicable" be of the same population size (*1*). From the strict premise of "one person, one vote" derives the constitutional requirement that the decennial census be as accurate as is practicable (*2*).

In addition, a wide array of federal funds—including multipurpose grants and financial assistance from programs running from education, housing, and social services through highway construction and energy conservation—is allocated on the basis of census figures (*3*). It follows then that the net undercount is important because it means votes and dollars.

Although census coverage has been good overall, census managers and customers alike have been troubled by the persistent differential undercount. Each decade the Census Bureau, and a new generation of census managers, has sought and received increased funding and has intensified efforts to reduce the differential undercount. Each decade census managers have expressed confidence in newly developed improvements in the science of census taking—for example, automation or better hiring and training practices. Each decade census managers have been optimistic, as the original enumeration (OE) got under way, that the undercount would nearly be eliminated. Yet each decade the differential undercount and the local variation in the undercount has proved unrelenting.

Conventional census methods offer little further opportunity for decreasing the differential undercount. In fact, increased effort may not reduce omissions but may actually increase double counting and other miscounts.

Why are people missed and miscounted? Census taking in 1990 was a massive and complex process that can be broken down into four basic steps: (i) compiling a list of addresses, known as the address control file (ACF); (ii) mailing forms out to each address so that the residents could fill them out and send them back; (iii) sending enumerators out to those addresses where the forms were not mailed back to get the census information; and (iv) a series of special activities, known as "coverage improvement programs" designed to find and count those people missed so far.

Census coverage errors can be sorted into two basic categories: omissions are people who should have been counted, but were not; and erroneous enumerations are counts that should not have occurred. Some important sources of these errors can be identified with each of the four steps listed above.

It is difficult to compile a complete list of addresses. In most areas, the Census Bureau starts with commercial address lists, usually used for marketing purposes. These lists are more complete for the more affluent neighborhoods. Although the bureau makes several attempts to update the address lists, there are certain types of housing that are hard to find. For example, when a house is subdivided into apartments, this may not be visible from the outside. Similarly, when people live in garages, backyard huts, and tents, these are often missed by enumerators, especially when they are located in poor neighborhoods with high crime rates.

When a residence is left off the address list, it is difficult for the people living there to be counted. Substantial numbers of omissions, perhaps 40 to 70%, occur because the entire housing unit is missed by the Census Bureau.

The next chance of omission occurs when residents fill out the form. If a family is nuclear with one or two parents and children, there is less of a problem in figuring out whom to include. On the other hand, Census Bureau studies have shown that the omission rates for distant and nonrelatives are very high. Those who move at about the same time as the enumeration also have a relatively high propensity to be missed. Literacy is another clear problem. When the census form is returned with certain persons missing, it will

Table 1. Estimated percent net undercount by race and sex since 1940 (*8, 15*).

Race and sex	1990*	1980	1970	1960	1950	1940
Total	1.8	1.2	2.7	3.1	4.1	5.4
Male	2.8	2.1	3.4	3.6	4.5	5.9
Female	0.9	0.3	2.0	2.6	3.8	5.0
Black	5.7	4.5	6.5	6.6	7.5	8.4
Male	8.0	6.7	18.6	8.3	9.1	10.3
Female	3.6	2.4	4.6	4.9	5.9	6.5
White/other	1.3	0.8	2.2	2.7	3.8	5.0
Male	2.0	1.7	2.8	3.0	4.0	5.2
Female	0.5	0.0	1.7	2.4	3.6	4.8

*Data for 1990 are preliminary. A number of modifications, expected to be relatively minor, may be made to these data during the next two years.

The author is vice president of the A. C. Nielsen Co., Nielsen Plaza, Northbrook, IL 60062-6288.

usually have no apparent errors, so there is no way of detecting the omissions and correcting them.

After about 6 weeks of mail returns, the census process shifts to the third stage, known as nonresponse follow-up. Instead of writing down the information, the resident gives it verbally to an enumerator. If it does not occur to the resident to include "Uncle Joe," you can be sure that the enumerator will not suggest it. Welfare mothers, those living in crowded conditions in violation of housing ordinances, among others, may feel reason to conceal residents. Such omissions are known as "within household omissions" and account for a substantial share of the total undercount.

So-called "coverage improvement programs" are designed to catch all who fall through the net, but history has shown that the programs are not fully effective (*4*). One of these programs, known as the "vacant delete check," sends enumerators to every address listed as vacant or not fit for habitation to make sure the classification is correct. The procedure is designed to add whole households and delete nonhabitable structures and does not reduce within household omissions. Other programs, such as the "parolees probationers check," are directed toward within-household omissions.

Erroneous enumerations include counts of people who died before or were born following census day, people who were enumerated more than once or in the wrong geographical location, and "people" who never existed but were nonetheless counted by an interviewer who simply created fictitious information in lieu of conducting a proper interview. Some people may have two residences and are counted at each. The Census Bureau devises procedures to try to catch such errors, but mistakes are inevitable.

Erroneous enumerations can occur as a result of any of the phases of census taking, but research has found that the later in the census process the counting takes place, the greater the likelihood of an erroneous enumeration. For example, in 1980, about 2% of persons listed on forms mailed back by residents were erroneous. Of the counts achieved later in various coverage improvement programs, 16% were erroneous. Early evaluations of the 1990 census display similar results.

Finally, I should note that omissions and erroneous enumerations sometimes occur together—as, for example, when census forms are stacked in a lobby and vandalized or when, in a small multiunit building, apartment A is missed and B is counted twice.

Methods of undercount measurement. Two principal methods of undercount measurement are in use for the 1990 census: demographic analysis (DA) and the post-enumeration survey (PES). DA is based on the demographic accounting identity

$$\text{Population} = \text{births} - \text{deaths} + \text{immigrants} - \text{emigrants}$$

and on the undercount identity

$$\text{Net undercount (\%)} = 100 \times (\text{total population} - \text{OE}) + \text{total population}$$

It was developed in the early 1950s by the Princeton demographer Ansley Coale and has been used subsequently by the Census Bureau to estimate the net undercount for each decennial census since 1940. With this method, the net undercount can be estimated for the nation as a whole by age, race, and sex. Difficulties arise in estimating local-area undercounts because there is no accurate means of accounting for internal migration between areas of the country.

The PES is based upon an intense sample survey conducted some time (usually months) after an original census enumeration. The data are arrayed in a two-way table (Fig. 1), displaying the fact that some people are counted by both systems (A), some by one or the other but not both (B and C), and some by neither.

The total population is estimated as the sum of A, B, and C plus some allowance for those people missed by both systems. The PES owes its origins to methods of estimating the size of wildlife populations and the underregistration of human births (*5*). It has been used by the Census Bureau several times, as early as 1950 and as late as 1990.

Fig. 1. Two-way table of post-enumeration survey (PES) and original enumeration (OE).

		PES	
		Counted	Missed
OE	Counted	A	B
	Missed	C	

During the 1980s, the Census Bureau invested around 100 person-years and $10 million in research, development, and testing of improved DA and PES methodologies for measuring undercount in the 1990 OE and of new methodologies for adding an allowance for the undercount to the OE so as to produce a final census count (a corrected count) that is closer to the true count than is the OE. A National Academy of Sciences (NAS) panel, along with many other advisory committees and academic consultants, participated in the research program.

The program ultimately produced the kind of census process needed to defeat the differential undercount at an affordable cost. This process, called the correction process, is based mainly on the PES approach with assistance by the DA method, and is currently being implemented by the Census Bureau for the 1990 census.

The correction process. I will summarize the correction process in eleven broad and nearly sequential steps.

Step. 1. An area probability sample of about 5000 blocks was selected. This sample size was thought to be sufficient to estimate the net undercount rate with a standard error of about 1.4 percentage points in each of 100 geographic areas. The sampling unit, the "block," was essentially a city block in urban and suburban areas, and a well-defined piece of geography in rural areas.

Step 2. The 5000 sample blocks generated two probability samples of people, the "population" or "P sample" and the "enumeration" or "E sample." The E sample consisted of all persons counted in the 1990 OE in those blocks and was used to estimate erroneous enumerations. The P sample consisted of all persons counted in an independent enumeration of the blocks conducted some time after the OE and was used to estimate omissions. The 1990 OE (and thus the E-sample enumeration) occurred mainly between late March and mid-June, whereas the P-sample enumeration took place mainly in July.

Step 3. As early as February, the Census Bureau sent fieldworkers to each of the 5000 blocks to create a list of every structure suitable for human habitation. This listing was completely independent of the way in which the address list was assembled for the OE. Each list misses a few addresses included in the other.

Step 4. The P-sample interviewing was conducted in person by Census Bureau fieldworkers who sought to count all residents of the sample block, including "in-movers" (that is, those who lived elsewhere during the OE). Because of undercounts and miscounts of various sorts, and because of differences between the in-movers and the "out-movers" (that is, those who moved from listed addresses between the original and P-sample enumerations), the lists of persons counted in the P and E samples fail to agree perfectly. The failure to agree provides a basis for use of both lists, in combination, to provide a more accurate count than either can provide individually. Approximately 160,000 housing units and 400,000 people were counted in the P-sample interview.

Step 5. The next several steps involved matching the P-sample persons to lists of persons counted in the OE. The match was based on name, address, and various demographic characteristics. The objective was to determine which P-sample people were counted in

the OE, and which were not. The initial phase of the overall matching operation was performed by an automated computer matcher, essentially an advanced expert system (*6, 7*).

Step 6. The computer matcher was able to match about 75% of the P-sample persons to their corresponding OE. All others, including all the in-movers, were referred to a trained team of clerks and professional statisticians who examined the information collected in both the OE and the P-sample interview and ultimately designated each P-sample person as matched, not matched, or "match status unknown."

Step 7. At this point in the process, each E-sample enumeration was either matched or not matched to a P-sample enumeration. Those E-sample enumerations that were matched were designated as "correct enumerations," meaning the corresponding person was correctly counted in the OE. All other E-sample enumerations were recontacted by fieldworkers for the purpose of collecting enough information to designate the enumeration as "correct" or "erroneous."

Concurrently, all P-sample persons who were designated as "match status unknown" were recontacted by the same fieldworkers for the purpose of collecting enough additional information to designate the person as enumerated or not enumerated in the OE. For example, the recontact may confirm that Mary Jane Peterson (P-sample enumeration) and Mary Jane Emerson (E-sample enumeration) are one in the same, with the surnames reflecting married and maiden names. At the conclusion of this step, almost all P-sample people were designated as enumerated in the OE or not enumerated, and almost all E-sample enumerations were designated as correct or erroneous.

Step 8. The data were then screened for any incomplete, missing, or faulty items. Typically, there is a small number of P- and E-sample people for whom one or two demographic items (age, sex, or race) are missing, and a small number for whom the enumeration status is still unclear. All of the missing data were completed by statistical imputation techniques.

Step 9. At this point in the overall process, the Census Bureau produced provisional estimates of true population size, and thus of percent net undercount (*8*). Estimates of total population were calculated within each of 1392 poststrata, based, in part, on the characteristics of the P- and E-sample people. One poststratum, for instance, consists of all black or Hispanic persons living in metropolitan statistical areas in the New England census division. The poststrata are mutually exclusive and jointly span the entire U.S. population. The estimator of total population within a poststratum, sometimes called the dual-system estimator, is of the form

$$\hat{N} = \frac{X - \hat{E}}{\hat{p}} \qquad (1)$$

where X denotes the actual population count achieved in the OE, $\hat{E}$ denotes an E-sample–based estimator of the total erroneous enumerations in the OE, and $\hat{p}$ denotes a P-sample–based estimator of the proportion of the total population that was enumerated in the OE. In other words, the numerator of $\hat{N}$ estimates the number of distinct, correctly enumerated persons in the OE, and this number is "projected up" by the proportion of the total population enumerated.

Another way of viewing $\hat{N}$ is by way of the two-way table (Fig. 1). Let $\hat{N}_{ij}$ denote a sample-based estimator of the total number of persons in the (i, j)th cell. The estimated total population may be expressed by

$$\hat{N} = (\hat{N}_{11} + \hat{N}_{12} + \hat{N}_{21}) + \theta\ \frac{\hat{N}_{12}\hat{N}_{21}}{\hat{N}_{11}} \qquad (2)$$

with $\theta = 1$. The first term on the right side in Eq. 2 is the total number correctly counted in either the OE or the P-sample, and this term alone tends to be a better count (that is, closer to the truth) than the OE. The second term on the right side in Eq. 2 is an estimator of the unknown cell (lower right cell in Fig. 1) in the two-way table, thus creating further improvement in the count.

Step 10. The 1392 ratios, $\hat{N}/C$, where C is the count from the OE, are called "correction factors." At this stage, however, the ratios are subject to high sampling variability, and further improvement may be achieved by "smoothing" the ratios, or in other words by reducing the sampling variability (*9, 10*). The "smoothed" correction factor is obtained by shrinking the factor toward a predicted value, with the degree of shrinkage determined by the quality of predictor and the inherent sampling variability in the direct factor.

Step 11. The Census Bureau will apply the correction factors to the OE, block by block, for each of the approximately 7 million blocks in the country. Conceptually, the process is one of correcting the pieces of each block, corresponding to the poststrata. This correction may be viewed as a process of developing undercount rates at aggregate levels, namely the poststrata, and carrying them down to local levels, namely the block. If the aggregate level corrected count is closer to truth, then so must be the local, corrected counts (*11*).

The counts are then rounded to an integer value with a controlled rounding algorithm. The controlled round will guarantee that the rounded numbers at one level of aggregation will add to the rounded number at a higher level of aggregation.

Step 12. At this point, the OE computer files will be corrected or completed for those originally missed or miscounted. Final 1990 census data products and tabulations may be prepared from these corrected files.

What can go wrong? Various errors affect the corrected and uncorrected counts. What matters is not whether either set is error free, but merely which is closer to truth. What matters is not whether the statistical assumptions underlying any part of the process are exactly right, but merely whether they are close enough to being right that they offer a basis for producing counts as accurate as is practicable.

One of the post-1980 research program's major contributions was the development of a taxonomy of PES errors (*12–14*), including eight major classes in the overall error structure. Census Bureau scientists developed the means of evaluating the sizes of the individual errors, as well as of the combined overall error. Such evaluations were incorporated in tests of the newly discovered correction process in 1986 and 1988. For example, in a 1988 test census in Missouri, the undercount was estimated to be 5.4%. Evaluations showed that this result was upward biased by approximately 1.0%, implying an estimated true undercount of around 4.4%.

The Census Bureau made relatively less progress during the post-1980 period on classification and analysis of error in the OE. Results giving the magnitude of error in each phase of the OE are needed. However, in each of the tests there is good reason to believe that the uncorrected OE was further from the truth than was the PES-based correction. For example, in a 1986 test census in Los Angeles, the PES-estimated undercount was 9%, and subsequent evaluation showed that the true undercount may be closer to 7.8%. Thus, the PES-based corrected count (that is, OE count × 1.09) was closer to the evaluation-based true count (that is, OE count × 1.078) than was the uncorrected OE count.

Turning to the 1990 census itself, it is important to emphasize that the Census Bureau designed its undercount measurement and correction process with built-in evaluation. As of this writing, there are 18 evaluation studies on the quality of the new PES correction

process as actually implemented. A number of critical studies will be completed by summer 1991, allowing a formal, scientific determination of which is the more accurate set of counts.

Next steps. Given the enormous difficulties, the 1990 OE was successful, but nevertheless encountered well-publicized difficulties, failing to achieve the levels of coverage hoped for. Across America, cooperation with the enumeration lagged behind benchmark levels, and ultimately only about 65% of the census forms were mailed back, reflecting a decline of about 10% from the 1980 enumeration. In turn, the lower mail-back rates passed a greater burden on to nonresponse follow-up and coverage improvement programs. Overall, many people were missed and many others were erroneously counted.

Meanwhile, during the last 18 months, the government published guidelines for assessing the quality and other features of corrected and uncorrected counts, and a special advisory panel was formed to advise the government on all relevant matters. As of this writing, we stand on the threshold of a decision—to be made by the Secretary of Commerce on or before 15 July 1991—as to whether to certify corrected or uncorrected counts as the official 1990 census.

Many groups and individuals have been commenting upon the government's guidelines and decision process. Overall, the process has been highly political. As might be expected, federal and local officials from both major parties have been examining the consequences of corrected versus uncorrected counts in terms of their own jobs, programs, and constituencies. Statistical scientists too have been commenting, with most of their emphasis on the accuracy and usability of 1990 census data products.

As one of eight members of the special advisory panel, I have followed the implementation of the 1990 OE and correction process from the beginning. Although I cannot speak for colleagues on the panel, I am confident—though not yet certain—that corrected counts will be closer than uncorrected counts to the truth. My confidence derives from two lines of thought. First, it now appears certain that, to a large extent, the PES-estimated undercount follows expected patterns, displaying higher undercounts where undercounts have been historically high or where difficulties were known to be especially severe in the OE operations. Second, by the very design of the correction process, corrected counts are guaranteed, in theory, to lie between uncorrected counts and the true population. This guarantee could fail in practice if the process had been implemented in slipshod fashion, with massive compromising errors. To the contrary, however, I am close to concluding that the Census Bureau's execution was far from slipshod, but rather was commendable and perhaps outstanding. It follows that corrected counts should be closer than uncorrected counts to the true population, both in absolute terms and, more critically, in the distribution of the population across states and other areas.

I will conclude my final assessment of accuracy over the next several weeks, using the Census Bureau's 18 evaluation studies, and I expect to be able to affirm my provisional confidence in the corrected data at that time. In any event, accuracy should be the basis for the secretary's decision on this matter.

REFERENCES

1. *Karcher* vs. *Duggart*, 462 U.S. 725, 730 (1983) [quoting *Wesberry* vs. *Sanders*, 376 U.S. 1, 7 (1964)].
2. *Carey vs. Klutznick*, 637 F. 2d 834, 839 (2d Cir. 1980).
3. General Accounting Office, "Grant formulas: A catalogue of federal aid to states and localities" (Washington, DC, 1982), pp. 384–399.
4. C. Citro and M. Cohen, *The Bicentennial Census: New Directions for Methodology in 1990* (National Academy Press, Washington, DC, 1990).
5. K. M. Wolter, *J. Am. Stat. Assoc.* **81**, 338 (1986).
6. M. Jaro, *ibid.* **84**, 414 (1989).
7. W. Winkler and Y. Thibaudeau, *An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Census* (SRD Technical Report, Bureau of the Census, Washington, DC, 1990).
8. U.S. Bureau of the Census, "Census Bureau releases refined 1990 coverage estimates from demographic analysis," press release of 18 June 1991 (CB91-222, Washington, DC, 1991).
9. E. P. Ericksen and J. B. Kadane, *J. Am. Stat. Assoc.* **80**, 98 (1985).
10. G. Diffendal, *Surv. Methodol.* **14**, 71 (1988).
11. K. M. Wolter and B. D. Causey, *J. Am. Stat. Assoc.* **86**, 278 (1991).
12. H. Hogan and K. Wolter, *Surv. Methodol.* **14**, 99 (1988).
13. M. H. Mulry and B. D. Spencer, *J. Am. Stat. Assoc.*, in press.
14. ———, *Surv. Methodol.* **14**, 241 (1988).
15. U.S. Bureau of the Census, "The coverage of population in the 1980 census" (Report PHC80-E4, U.S. Government Printing Office, Washington, DC, 1988).