sus never intended such an interpretation. In the 1970 Swedish census, Statistics Sweden presented two numbers: one regular set of estimates with missing data and one with imputed values added. Surprisingly, many users (but perhaps not so surprising after all) knew exactly which estimate to use. In the 1975 census, imputation was not performed, which made comparison to the 1970 census awkward.

We find it both reasonable and natural to use auxiliary information to improve an estimate. After all, this is what survey design is all about. Various model assumptions are made in every design step, but the final result should be expressed as a single count or estimate. We sympathize with Belin and Rolph regarding their general conclusion about the protracted controversy on the undercount problem. An impressive amount of work has been done, but it appears as if we have reached the point where further methodological resources, time and money would be a waste.

Most U.S. statistical agencies have committed themselves to modern quality thinking, that is, various forms of total quality management. It seems as if it would be better to use the "debate resources" to improve the regular census count procedures, thus decreasing the need for extensive and expensive evaluation procedures. This is especially true for the U.S. undercount, where the discussion fails to result in a consensus.

Comment

David Steel

Evaluating and possibly adjusting the census for undercount raises a lot of difficult statistical and general issues. The papers here consider several of these and add to the already large literature on the subject. While the basic questions are now clear, the answers are not. To enable readers to make a judgement about any prejudices I might have on these issues, I should point out that as a former officer of the Australian Bureau of Statistics (ABS) I was involved in the evaluation of the 1981 and 1986 censuses. While the views I have are entirely my own, they are influenced by this past involvement. In terms of my prejudices this could work either way: having been involved in adjustment, I may have a bias to that view to justify my past work; alternatively, detailed knowledge of the many problems involved could lead me to be against adjustment.

In Australia, population estimates based on census counts adjusted for undercount have been released as the official population estimates since 1976. The estimates are produced for states and territories and local government areas. Population estimates are used to determine the number of seats each state has in the federal House of Representatives and the allocation of funds to states and local government areas. The decision to adjust was prompted by the high undercount rate showed by the 1976 Post Enumeration Survey (PES) and the fact that the 1976 census count fell considerably below the population estimates for 1976, which were based on updated 1971 census results. There has been general acceptance and remarkably little controversy surrounding the adjustment. A clear distinction is made between census counts and population estimates. Census counts are produced without any adjustment. There are similarities to the situation in the United States. The level of undercount is basically estimated from a PES which involves an independent household survey and matching between the census and the survey to determine missed people and some categories of erroneous enumerations. Dual system estimates (DSE's) are calculated. The results of the PES are compared with demographic analysis and other population indicators such as school enrolments and Medicare enrolments (Medicare applies to all age groups), primarily at the national level, but with some analysis below this. Synthetic estimation is used to obtain population estimates for local government areas. The procedures for the PES and census evaluation are decided in advance. The view is that quality must be designed into the census and the PES. The estimated level of net undercount is remarkably similar to the United States: 1.9% in 1986 and 1.8% in 1991. In 1991 the state undercount rates ranged from 1.2 to 4.1%. The ranking of the states in terms of undercount has been consistent over time. Further details are given in Choi, Steel and Skinner (1988), Trickett (1992) and Australian Bureau of Statistics (1990).

David Steel is Senior Lecturer, Department of Applied Statistics, University of Wollongong, Northfields Avenue, Wollongong, New South Wales 2522, Australia.

There are also some important differences between the situation in Australia and the United States. The Australian census is conducted by field methods, and collection of census forms is essentially completed in a two-week period, enabling a PES to be conducted approximately three weeks after the census (census forms received after the PES has started are excluded from the DSE's). Conducting the PES as close as possible to the census date is an important factor in improving the quality of the census evaluation. The PES includes about 40,000 households comprising 93,000 people but is much less clustered than in the United States. Over 4,000 clusters are selected, giving a good geographic spread. This can be done since there is no direct equivalent to the E-sample. Attempts to match people selected in the PES sample are made with census forms for the selected dwelling and those in the surrounding areas and at other addresses provided at the PES interview where people may have been enumerated (e.g., usual residence of visitors and the surrounding areas). This provides information on the major categories of erroneous enumerations. Some classes of erroneous enumerations will be missed, such as households which are completely fabricated in the census. In the matching process unresolved cases are finally imputed using a regression-based method.

This size of the PES sample and its design means that direct estimates can be calculated for the eight states and two territories. For each state, separate estimates are calculated for the capital city and the rest of the state. This leads to estimates for 14 major geographic areas, referred to as parts of state. Estimates for local government areas are then obtained using synthetic estimates using age–sex poststrata within each part of state. This means that a lot of the concern about using undercount rates for other states is avoided. However, there will be variation in the undercount rate that is not accounted for by using the age–sex poststrata. The approach has been deliberately conservative, adjusting for factors which have clearly been demonstrated to affect the undercount rate. It is felt that such an approach will bring the population estimates closer to the true distribution, but will probably understate the variation in undercount rates. Hence, while it should lead to improvements, the danger of overadjusting is reduced. Other methods of producing undercount estimates for local areas have been investigated (Steel and Poulton, 1988; Bell, Cornish, Evans and Vincente, 1993) but the current synthetic procedure is considered the most appropriate.

There is little evidence in Australia of difference in undercount rate between different ethnic groups in the community. The exception is the Aboriginal community, whose level of undercount is, to some degree, associated with problems of enumeration in remote areas. Analysis of undercount rates for different groups by birthplace has shown little differences from the Australian born. The group with the highest undercount consists of those born in New Zealand, which is probably due to the relative youthfulness and mobility of this group in Australia. Analysis has shown that the undercount is high for those who were away from their usual residence on census night, with an estimated undercount rate of 16.5% in the 1991 census (Trikett, 1992). Matching procedures for this group are especially important. In 1991 the scope of the areas checked for such people in the matching was expanded, with beneficial results.

Demographic analysis, using the 1921 census as a base, is used to validate the PES results and on occasion make some adjustments to these figures at the national level, which then flow through to lower levels. Reliable birth and death registration systems and a system to measure overseas arrivals and departures have been in operation for a long time. Movements in and out of the country are relatively easy to monitor, with there being limited points of entry. Five yearly gaps in the census also help in the demographic analysis. The view has been taken that the PES is generally more reliable than the demographic estimates, which are mainly used to evaluate the PES. The PES has some problems. A small number of blocks can have undue influence, conducting independent check in remote areas is difficult, matching is not perfect and erroneous nonmatches can inflate the estimated undercount rate. In 1991 there was a downward revision in the estimated resident population for Western Australia. It is thought that the 1986 census count had been overadjusted because of the effect of school holidays on the number of persons away from home at the time of the census (Trickett, 1992). However, based on the evidence and knowledge of the processes used, the judgement has been that population estimates incorporating an adjustment will give a better estimate of total population and population shares than unadjusted census counts. This is a judgement which is finally made by the Australian statistician, whose independence is guaranteed by the ABS Act of 1975. I would expect any concerns about the approaches used would be raised and handled through various mechanisms that exist for state government input into ABS activities, not through the courts.

The Australian experience gives an example where reliable estimates of census undercount can be made and where adjustment is carried out. However, there are sufficient differences not to see this as necessarily endorsing the entire procedure originally proposed for adjusting the U.S. census. The paper by Breiman raises the fundamental issue of what is the

quality of the PES. This a good question to raise and is logically the first to ask. To answer this question involves making an assessment of nonsampling errors, and it is notoriously difficult to get a complete picture of these. The concern is whether the quality of the evaluation is sufficient for its purposes. This is exactly what Breiman is raising concerning the undercount estimates, but the same question can be raised about the quality evaluation of the PES that he reviews and summarises. There are many ways in which a PES could potentially give bad data. While the paper raises some doubts, the case is made in a way that, despite the author's attempt to bring a lot of detailed evidence together, I have no feel for what the errors that might be there would do. The quoting of percentages on what appears to be different bases makes it hard to work through what the likely effect might be. Breiman notes that two 0.5% differences, working in different directions, can affect the estimated undercount rate by 50%. Essentially this point is that, to first order, if 1% of the PES is not matched when it could have been, then the estimated undercount rate will increase by approximately 1 percentage point. Some of the rates that are quoted are calculated on different bases, which may be relevant, but to asses their impact on the undercount estimates the reader needs them expressed as percentages of the entire PES sample. Quoting gross differences also makes independent assessment difficult. A tree diagram of what happen to the PES sample in terms of matching cases and nonresponse would make interpretation easier.

The paper by Belin and Rolph offers some different interpretations. Again we have reasonable statisticians offering a significantly different interpretation of the same data. To help in understanding the situation I look for other evidence. With the adjustment for the processing error mentioned by Breiman the estimated undercount is 1.7%, which is close to the demographic estimate of 1.85% provided by Robinson, Ahmed, Das Gupta and Woodrow (1993). This does not suggest gross overestimation of the undercount rate through the PES. Both estimates are imperfect but their consistency offers some reassurance. Of course it can also be argued that both systems have errors that just happen to give similar results.

The paper by Freedman and Wachter raises valid questions about synthetic estimation. Synthetic estimates are biased; so are census counts. No adjustment is a model. Hence it is desirable to incorporate an allowance for these biases for both estimates. The paper sheds more light on this issue. One result of the analysis is that, based on the proxy variables analysed, the synthetic estimates do explain much of the variation in rates between states. For example, using the root mean squared error in Table 6 and comparing it with the standard deviation across states in Table 5 for the substitution variable, we see a 42% reduction in the variance of the error of prediction when using the synthetic estimates as against the national rate. Applying the national rate would be the same as no adjustment in terms of population shares.

I have sympathy with the comment of Wachter, quoted by Belin and Rolph, on the use of aggregated measure. I do not think this precludes using empirical evidence to resolve issues, but implies looking beyond summary measures. I am therefore surprised that Freedman and Wachter do not provide any plots of the errors or bias of the synthetic estimates. It would have been interesting to see the distribution of the errors and any relationship with the size of states or other factors.

To adjust the census to provide population estimates, the reliability of the adjustment must be better than the error it is trying to correct. There are several ways of assessing the quality of the PES and the associated estimates, none of which is entirely complete or satisfactory, but which in combination should enable a judgement to be made. The process of assessing the quality of the PES has involved several sources. One source is the type of evaluation studies reviewed and summarised, with different conclusions, by Breiman and by Belin and Rolph. These studies themselves will have quality problems and are not complete (but I do not think we need to evaluate the evaluation of the census evaluation—we must stop somewhere). The quality that has been designed into both the census and the PES and the subsequent processing and the quality assurance procedures are factors which decision makers must take into account. Consistency with past results, or at least reasonable explanations of the differences, are also indicators. Comparison with other sources, such as demographic estimates add to the picture. Finally, what is sometimes called face validity, or consistency with what is thought to be known about undercount, is a factor. Some of these factors are "hard" and some "soft," but the overall judgement should be made by combining these assessments together. In Australia that has led to adjustment for the purposes of population estimates. To answer the question raised by Belin and Rolph, I do not think a consensus is possible. The standards of proof required by some are just not achievable. On what major issue would the profession have a complete consensus? Decisions have to be made on imperfect, sometimes contradictory evidence, in the face of strong disagreements by reasonable, well-qualified and well-intentioned professionals.