DISCUSSION OF SOME CENSUS 2000 DRESS REHEARSAL EVALUATIONS

Mary H. Mulry, M/A/R/C Research
Mary H. Mulry, M/A/R/C Research, 7850 North Belt Line Road, Irving, TX 75063

Key Words: Master Address File, Post-Enumeration Survey, coverage, undercount

First, I want to say that I am very pleased to be invited to discuss papers on the evaluations of the 1998 Dress Rehearsal. I want to applaud the Census Bureau for its desire to improve and refine the census and the census evaluation process.

These papers consider very important aspects of the 2000 Census. We have two papers on the Decennial Master Address File (DMAF), one on the approach to the quality control for Post Enumeration Survey (PES) interviewing, and one on the error profile for the PES.

These represent fundamental components of the 2000 Census. Remember that there will be two sets of census numbers, one based on the census count and another set adjusted with results of the Accuracy and Coverage Evaluation (ACE) survey that uses PES methodology. The ones adjusted with the results of the ACE will be available for use in Congressional redistricting, drawing districts for state legislatures, fund allocation, and as the basis for controls in many government statistics.

First I am going to discuss the paper "Consistency of Housing Unit Data with Demographic Benchmarks" by West and Robinson.

The idea presented in the paper has the potential to be a very good tool to identify problem areas in the DMAF. I do have a few comments. The paper contains estimates from the demographic benchmark, the DMAF, the 1990 Census, and the 1998 Dress Rehearsal.

One question that strikes me is 'Which housing unit count do I believe?'

Obviously, no one believes the 1990 Census. The authors expect the DMAF estimates to be high, although they will be revising their estimate of the expected percentage error. The remaining choices are the demographic estimate and the Dress Rehearsal count. Sometimes the DMAF is closer to the Dress Rehearsal count than the demographic estimate.

I would like to see a comparison with an estimate of the housing unit coverage from the Dress Rehearsal Post Enumeration Survey. These estimates are a by-product of the new matching operation. First the housing units are matched and then the people are matched. This is one of the improvements in the design that has occurred during the 1990's. The 1990 PES did not include housing unit matching. The 1990 Housing Unit Coverage Study was done long after the undercount estimates for people were processed and calculated. Since these data are available from the 1998 Dress Rehearsal, designing how to use the demographic housing unit counts is a perfect way to use them.

Another question is 'What is a large discrepancy?'

We should have different expectations for tolerances on the percentage discrepancy for different levels of geography. For smaller areas, such as census tracts or some counties, the threshold for the discrepancy to cause concern would be relatively high. For larger areas, such as those the size of the Dress Rehearsal sites and larger, the threshold would be relatively low.

When looking at the discrepancies for tracts, 10% is relatively low. However, a 10% discrepancy for a site would be very large. My challenge to the authors is to design thresholds that are appropriate for the different levels of geography.

A good resource to use when developing the thresholds is the housing unit coverage estimates. The PES housing unit estimates probably are the best set of estimates for the Dress Rehearsal sites. Comparison of the DMAF, demographic estimates, and 1998 Dress Rehearsal housing unit counts with the evaluation of the housing unit coverage will provide insight on how to set the thresholds.

Another approach to explore when setting thresholds is the use of a regression model to predict the discrepancy for a geographic area. The discrepancies seem to the related to the size of the geographic area, the nature of the geographic area, the estimated growth, and the Hard-To-Enumerate score. These all could be used as explanatory variables in a regression model that predicts the percentage error. The predicted percentage discrepancy would be compared to the observed discrepancy.

Next I will turn to the "Quality Improvement Program for the DMAF" by Burcham and Barrett.

The size of the gross errors measured in the evaluation appears large. Both the percentage erroneous inclusions and percentage missed are in the 10% range. When we consider that the undercount in 1990 was 1.6%, these appear to be large numbers. Certainly we would like to see a more accurate DMAF. The more accurate the address list, the better the census has to be.

However, I find the amount of error hard to judge. The DMAF concept of merging the US Post Office Delivery Sequence File and the previous census address list is a new way of forming an address list for the census. In 1990, the address lists for urban areas were purchased from vendors. The Census Bureau created the lists for the suburban areas and the rural areas where the census forms were mailed.

Data from the 1990 Census that evaluates the quality of the initial address lists would be informative. Such a comparison would set some context for evaluating the DMAF errors.

There really are two issues for the DMAF: coverage of housing units and coverage of addresses. Even if the DMAF had every single housing unit in the nation, but a substantial percentage of the addresses were incorrect, that would be a major problem. The Post Office will not be able to deliver the census forms correctly if there are errors in the addresses. They can compensate only so much for address errors.

The paper contains a disclaimer that the evaluation may not accurately portray the gross error with respect to the inclusion of housing units. The reason is that some of the interviewers may have reported on the accuracy of addresses, not whether the housing unit was included.

Keep in mind that incorrect delivery of census forms was a problem in the 1990 census. There is no reason to think that we are immune to that problem in Census 2000. Having accurate addresses is one way to minimize the mail delivery problems.

Therefore, we need to know about both: existence of housing units and existence of addresses.

I do think that analyzing the data with and without the unresolved cases is a good way to go to study the robustness of the estimates. However, in this situation, an unresolved case is an indicator of a potential problem. Not being able to tell whether the HU or address exists is an indicator of a problem for an NRFU interviewer who may have to sort out which housing units to interview. So, I do think that including the unresolved cases with the assumption of the worse case scenario for them is not an exaggeration of the errors.

Now I will address the paper "Evaluation of the Interviewer Quality Assurance Falsification Model" by Krejsa, Davis, and Hill.

In my opinion, the quality control (QC) of the ACE in 2000 is a very important operation. The quality control of the 1990 PES was a major contributor to its success. I think the targeting of interviews that appear out of bounds is a very good addition to a random sample. This is definitely an improvement over the random sample that was part of the 1990 methodology.

In 1990, if an interview failed the QC, the interviewer's entire work unit (the interviews for that same day) was re-interviewed. However, I did not see any mention of such a procedure in this paper. I hope this procedure is still in place and was just not mentioned.

I realize that the study did not produce very much data, but I still would like to see whether any additional interviews were fabricated when a falsification was discovered in the targeted or random sample. Such a calculation would give us more insight into the effectiveness of targeting.

I think that the variables used in the targeting are good choices. Possibly combinations of variables would be even more effective than using them one at a time. However, I was surprised to see that they were evaluated one at time. I suggest exploring synthesizing these variables by making them explanatory variables in a logistic regression model that predicts the probability that the interview is a potential falsification.

I do have a comment on the topics under discussion. I would not stop the QC of an interviewer after the interviewer has passed an initial QC period. An initial QC would detect those that do not intend to do a good job. However, there are those who usually do their job, but may falsify interviews when they get in a case that is difficult to contact. They try to do an interview once or twice, but do not go back multiple times for whatever reason, and then falsify the interview.

The last paper is "Error Profile for the PES in the Census 2000 Dress Rehearsal" by Bean, Bench, Davis, Hill, Krejsa, and Raglin.

This paper is really more like four papers in one. I am not really sure what to say about this paper. I find myself perplexed. I do not see what I would like to see in it, and I am undecided about what to say about what I do see.

I think that assessing the error for an application of PES methodology is important. The reason is that PES

methodology is complicated and requires some vigilance to be sure that it is implemented correctly.

In these evaluations, I find assessing the impact of the matching error and data collection error to be difficult. I would much rather see a total error model. Since the ACE estimates will be available for redistricting and fund allocation, an assessment of the error that confirms their validity will be important. By not having a total error application in the 1995 Census Test and the 1998 Dress Rehearsal, the Census Bureau has missed valuable opportunities to refine and adapt the methodology to the current processing.

The paper does provide information about several sources of potential error: matching error, data collection error, and mode collection error.

As for the Matching Error Study, the low level of error measured does not surprise me. The tests of PES methodology always show low matching error. We found that to be the case leading up to the 1990 Census.

What is important for the Dress Rehearsal is that the matching error study be done. The focus is the preparation for assessing the matching error in the 2000 ACE.

I do think that the gross difference rate (GDR) is the appropriate measure to use in evaluating the Dress Rehearsal matching. What I did not see in the tables was movement from Unresolved to Resolved or Out-of-Scope to Resolved. I always found the movement between the status of Unresolved and Resolved to be very instructive. There are some cases that are on the edge of the classification. I prefer to see a comparable number switching each way.

As for the Evaluation Follow-up, I do not think that the GDRs are the appropriate measure to produce with the data. The study was designed to identify the type of data collection error that caused an error, but changes in the study were necessary. Even though the implementation did not go as planned, additional information was collected about the sample cases. So matching using all the information should be better than the original process. Using the results as a standard of comparison would provide insight into the error and the robustness of the estimate.

The GDRs shown appear very large in some cases. My question is whether there is any explanation for this. For example, 'Did these cluster by household or by block?' or 'Were they caused by cases with big weights?'

Unfortunately, no matches were sent to EFU. Plans called for sending a sample of matches to EFU, but this part of the operation was canceled. The 1990 EFU did find matches that were really non-matches.

As for the evaluation of the decision to not send non-matching people to follow-up when their housing unit matched, the results show there are few errors. In 1990, we sent this type of nonmatch to follow-up, but we did not have housing unit matching when we were matching the people. If the housing unit matches, but the people do not, then assuming that they are census misses rarely will be an error. However, I do suggest that a sample of these cases be sent to follow-up, just as a sample of matches are sent to follow-up.

The last study is the data collection mode evaluation. The study compares collecting PES data on the phone from early census returns with in-person interviews. Interviewers collected some of the 1990 PES and EFU data by phone. When the interviewer could not get an interview in person, sometimes they used the phone. In the EFU, when the interviewers in urban areas found that a PES respondent had moved out of the PES sample block, they often checked directory assistance for a new phone number, and conducted the interview by phone when a new number was available. There are no records of the number of phone interviews.

The 1990 EFU experience leads me not to be surprised that there are no mode effects. The sample sizes are low, but I doubt we would see differences if they were higher.

What we do need to consider for this operation is the indirect consequences. There may be some other implications for the data beyond the individual cases alone. Some questions I have are:

1. What are the implications of the interviewers not visiting these housing units in person?
2. Did having fewer cases to visit make the interviewer's job easier or faster?
3. Was there any advantage in the matching process?
4. Did the early initial returns cluster by block? There are some blocks that match completely, and the early returns may cluster in blocks that tend to have a high match rate.
5. Does this phone operation expedite the easy blocks? If so, that would leave more resources for the hard blocks.

The Census Bureau faces a huge challenge next year in processing a census and an ACE. I am glad that the plans include an evaluation of the ACE. I look forward to seeing the result.