Bias in the Post-Enumeration Survey due to Duplicate Misclassification

Scott Konicki[1]
United States Census Bureau, 4600 Silver Hill Rd, Suitland, MD, 20746

Abstract
The 2020 Post-Enumeration Survey used dual-system estimation to estimate the net coverage of the 2020 Census in the United States and Puerto Rico. Part of the dual-system methodology involves selecting a sample of census enumerations and assessing whether the enumerations are correct or erroneous. A source of erroneous enumerations in the decennial census is duplicate enumerations, in which two or more census records correspond to the same unique person. When one record of a duplicate pair is selected into the Post-Enumeration Survey sample, we expect that we are equally as likely to have sampled the correct record of the duplicate pair as we are to have selected the erroneous record. In this paper, we use data from the 2020 Post-Enumeration Survey to show that the match codes assigned by clerical matchers are biased towards calling the in-sample record the correct enumeration, which results in an upward bias of the dual-system estimates. We explore the bias caused by this misclassification for the 2020 Post-Enumerations Survey and discuss the implications for future iterations of the survey.

Key Words: post-enumeration survey, coverage error, duplication

1. Introduction

The 2020 Post-Enumeration Survey (PES) measured the coverage of the 2020 Census by producing estimates of net coverage error and the components of coverage. To estimate net coverage error, the 2020 PES estimated the number of people in the population using dual-system estimation and compared this estimate to the census count. The dual-system estimation required two independent systems. The Population (P) sample was a sample of the household population that was independent of the 2020 Census. The Enumeration (E) sample was a sample of census enumerations in the same sample areas as the P sample. The P sample provided information about the population missed in the census, and the E sample provided information about erroneous census inclusions. Marra and Kennel (2022) provide more information about the source of the 2020 PES data.

This paper is concerned with duplicate census records identified by the E sample. The 2020 PES estimated that the 2020 Census included 7.2 million (standard error 0.1 million) erroneous person enumerations (Khubba et al. 2022). Of these, 5.2 million (standard error 0.1 million) were due to duplication. To estimate duplicates, we matched the E sample to the census to identify duplicate pairs, i.e., two census records that corresponded to the same

[1] Any opinions and conclusions expressed herein are those of the author and do not reflect the views of the U.S. Census Bureau. The U.S. Census Bureau reviewed this data product for unauthorized disclosure of confidential information and approved the disclosure avoidance practices applied to this release. DRB Numbers: CBDRB-FY22-136, CBDRB-FY22-DSEP-001, CBDRB-FY22-332, and CBDRB-FY22-418.

unique person. For a duplicate pair, only one of the records can be a correct enumeration. Thus, the PES must determine for every duplicate pair which record is the correct enumeration, called the primary record, and which record is the erroneous enumeration, called the duplicate record. Information from the PES interview, matching, followup interview, and clerical review was used to make this determination. The primary record is where the person should have been counted, according to the PES assessment of the situation in accordance with the census residence rules.

Some duplicate pairs had both records in the same census block.[2] In this situation, the E sample generally included both the correct and the erroneous enumeration. For duplicate pairs in which each the two enumerations existed in different census blocks, the E sample was likely to include only one of the records. Across all possible realizations of the PES sample design that include this duplicate pair, the E sample was, roughly speaking, equally as likely to include the primary record as it was to include the duplicate record. This is because the sample design is in not informed by the duplicate pairs and the assignment of the primary record. Thus, we would expect that for all records in the E sample that are part of duplicate pair, about half of these records would be determined to be the primary record and half would be determined to be the duplicate record.

For the 2020 PES, we found that this was not the case. Rather, for the set of records in the E sample that were part of a duplicate pair, we found that the sample case was more often called the correct enumeration. This imbalance leads to an overstatement of correct enumerations and thus an upward bias in the dual-system estimate (DSE). As noted later, this is not an entirely new problem, as the post-enumeration survey for the 2000 Census also overstated the correct enumerations for duplicate pairs (U.S. Census Bureau 2004).

This paper examines the magnitude of this bias and explores an adjustment to the E sample to correct for the bias. Using this adjustment, we recalculate the DSEs and compare the new net coverage results to the production results in Khubba et al. (2022). This paper also discusses reasons why this error may have occurred and further implications for the 2020 PES estimates of net census coverage error.

All comparative statements in this report have undergone statistical testing, and, unless otherwise noted, all comparisons are statistically significant at the 10 percent significance level.

2. Motivation and Background

2.1 2020 Census Coverage for People Aged 18 to 29

The Census Bureau uses two principal methods to evaluate the coverage of the decennial census. The PES is the survey-based method that uses dual-system estimation to estimate the size of the population. The second method is Demographic Analysis (DA), which uses vital records and other sources to create an estimate of the total population. In a Census Bureau blog, Jensen and Kennel (2022) compared the PES and DA estimates of net coverage error for the 2020 Census. As noted in this blog, the PES and DA estimates showed similar patterns of net coverage error. One age group with a notable difference was the population aged 18 to 29. For this group, the PES showed statistically significant

[2] More specifically, the same basic collection unit, or BCU. The BCU was the smallest geographic level for 2020 Census data collection and roughly corresponded to a block. The 2020 PES sample was a sample of BCUs. In this document, we use the term block for simplicity.

undercounts for both males and females while the DA estimates generally showed overcounts.

The blog explains that this age group includes many people living in group quarters facilities, especially college dorms, which are out of scope for the PES but in scope for DA. This difference in universes for the PES and DA may explain some of the difference in the net coverage error results. For example, if the 2020 Census overcounted the people aged 18 to 29 who were living in group quarters, this would not be reflected in the PES coverage estimates but would push the total DA estimates towards an overcount. Another explanation for the difference in the PES and DA coverage estimates is that one or both of the programs had errors in the estimates for this age group. I was interested in whether the PES was overstating the household population (including students living in off-campus housing units) for this group, and thus estimating net undercounts.

There could be many reasons why the PES may have overestimated the population of young adults. We know that the COVID-19 pandemic caused many challenges with counting college students. Many colleges closed in the spring of 2020, and students returned home to finish the semester virtually. The census reference day was April 1, 2020, so these students were often living at home during the time of the initial census data collection. Many students remained at home through the fall semester and later, during which time the PES independent interviews were conducted. The census residence rules state that college students who are living away form their parents' or guardians' home while attending college are to be counted at the on- or off-campus residence (U.S. Census Bureau, 2018), even if they moved elsewhere before April 1, 2020 in response to the pandemic. The pandemic-related disruptions to the living situations of college students may have exacerbated confusion over where to count college students in the census and whether to include them as in scope for the PES, and this may have led to an overestimate of this population by the PES.

Determining whether and to what extent the PES made these types of errors is not straightforward. Instead, I investigated a possible overstatement of the population by the PES by drawing on an important property of duplicate census enumerations, as explained later in this paper.

2.2 The Dual-System Estimate

The U.S. Census Bureau formulates the DSE as follows (Wolter, 1986):

$$DSE \approx N_{+1}\left(\frac{N_{1+}}{N_{11}}\right)$$

Here, (N_{1+}/N_{11}) is the inverse of the match rate for the PES, and N_{+1} is the number of correctly enumerated people in the census. There are two general ways in which the PES can overestimate a population total. The first is by understating the match rate. If the PES fails to identify matches between the P sample and the census or overrepresents the population missed by the census (relative to the population counted in the census), then the estimated match rate will be too low, and this would increase the DSE. We generally don't think this is an issue. In fact, previous post-enumeration surveys have included adjustments to increase the DSEs of adult males because the P sample was suspected to have overrepresented people who were counted in the census, relative to those who were missed (Shores 2002; Konicki 2012).

The second way the DSE can overestimate the population total is by overstating the number of census correct enumerations. If the PES fails to identify erroneous enumerations, such as duplicates, then the correct enumeration term and the DSE will be too high. This was the issue with the initial coverage results for the 2000 Census, causing the Census Bureau to conduct extensive research into the quality of the 2000 post-enumeration survey and issue revised estimates. The 2000 post-enumeration survey initially estimated a net undercount of 1.18 percent, but the revised estimate was a net overcount of 0.49 percent. Refer to U.S. Census Bureau (2004) for more information about the 2000 PES (called the Accuracy and Coverage Evaluation).

2.3 Duplicates in the Census

For this paper, I examined whether the 2020 PES potentially overstated correct enumeration by looking at duplicate pairs identified by the E sample. Before discussing the details, I first provide some examples of how duplication can occur in the census.

First, consider the Doe family, consisting of a married couple and their daughter Jackie. The Doe family correctly responded to the census at their home address using the internet in March 2020. Across town, there was a vacant house. Because nobody lived here at the time, there was no census response for this address. Meanwhile, suppose that in June the Doe family moved into this address. Later, say in August, a census enumerator visited the address to enumerate the nonresponding unit. This interview should have been about who, if anyone, lived at the address on April 1 (the reference day, called Census Day). However, either the census enumerator did not make this clear or the Doe family misinterpreted the request. Instead, the Doe family kindly supplied their information again at this new address, and now they have been counted twice in the census. The whole household has been duplicated.

As a second example, again consider the self-responding Doe family. In this example, suppose that the daughter Jackie was a college student who lived away from home during the semester at an off-campus apartment. The Doe parents erroneously included Jackie on their home census response, thinking that she should be counted as part of their family. Meanwhile, Jackie and her roommates at the off-campus apartment correctly responded to the census by counting themselves at this location. Thus, Jackie has been counted twice.

2.4 Identifying and Classifying Duplicate Pairs in the PES

The PES identifies duplicates by matching the E sample (the sample of census enumerations) to the entire census. When two census records are linked and believed to represent the same unique individual, we call the two records a duplicate pair.[3] This matching is aided by information from the PES independent interview in which we ask for other places people may have been counted in order to help narrow the search area.

For each duplicate pair, the PES must then make a determination as to which half of the duplicate pair is the primary enumeration and which is the duplicate enumeration. The primary enumeration is the correct record, which is where the person should have been counted. The duplicate is the erroneous record. Only one of the records can be correct, but

[3] In this paper, I discuss duplicates as if there are two census records for an individual. There are rare instances where three or more census records were found to correspond to the same individual. Since this was rare, I ignored these cases in much of my analysis and anticipate that this had a little impact on the results.

unfortunately this determination is not always easy. When the two records are not in the same census block, it is likely that only one of the records will be in the PES sample, meaning that we will only have the additional PES interview information for that one record. A couple things could go wrong with the primary and duplicate classification.

First, PES respondents may have trouble remembering where they lived on Census Day. Consider the example in which the Doe family moved in June. When interviewed by the PES in late 2020, they may have forgotten when they moved and erroneously reported that they were living at the new address in April. We call this recall bias. In this instance, the PES would incorrectly classify the census enumerations at the new address as the correct halves of those duplicate pairs.

A second thing that could go wrong is that regardless of which address is in sample, the PES respondent may continue to report the duplicated person as living there. Each census response included the person, so it is reasonable to believe the PES response at either address would also include the person. The respondents may remember their responses to the census and wish to provide a consistent report, or otherwise continue to report who they believe is living at the address. Consider the example in which the daughter Jackie was duplicated. If the parents' home is in the PES sample, then the parents could continue to insist that Jackie is part of the family living here, that she is only away temporarily. Similarly, if the off-campus apartment is in sample, then the roommates may continue to report that Jackie in fact lives here most of the time. If these reports are taken at face value without a careful consideration of the residence rules, then the PES would mark the sampled enumeration as correct regardless of which case is in sample. However, in truth, only one of these can be correct, and that determination should not depend on which half was selected in the sample.

A third possibility is that the clerical matchers may be biased towards assigning correct enumerations, especially if other people in the household have been determined to be correct. It may feel good to call a person a correct enumeration even if there is some uncertainty about this determination. This confirmation bias would make Jackie be a correct enumeration regardless of which address is in sample. Again, consider the example in which the parents' home is in sample. There is no uncertainty or duplicate link for the parents, so they are unquestionably determined to be correct enumerations. Jackie has some uncertainty because of the duplicate link to the off-campus apartment. A clerical matcher may see that the parents are correct and think that this must be where Jackie lives as well, thus incorrectly marking her as a correct enumeration at this address.

These examples hint at a helpful property of duplicates pairs in the PES for identifying whether misclassification of the primary and duplicate enumerations occurred. The PES sample is wholly independent of the primary and duplicate determination. Loosely speaking, this means that for any given duplicate pair, the PES is equally likely to sample the primary enumeration or the duplicate enumeration. This isn't exactly true because the PES has a complex sample deign and the two census records may be in different sampling strata with different probabilities of selection. However, this independence does mean that for all records in the PES sample that are part of a duplicate pair, we expect that the weighted sum of those cases which are the primary enumeration will equal the weighted sum of those cases which are the duplicate enumeration. That is, there should be about a fifty-fifty split of primary and duplicate enumerations in the PES sample. If there is a departure from this fifty-fifty split (beyond that which can be attributable to sampling

error), then this would be an indication of misclassification of the duplicate pairs for some of the sample enumerations.

The aggregate misclassification can occur on either side of the fifty-fifty split and the implications are as follows. If the sample cases are more often called the primary enumerations, then the correct enumeration rate as estimated by the PES will be too high. This will lead to a higher dual-system estimate, thus overstating the total population and pushing the net coverage estimates towards an undercount. The opposite is true if the sample cases are more often called the duplicate enumerations, or if the PES failed to identify the duplicate pairs in the first place. Here, the estimated correct enumeration rate will be too low, the population is understated, and the net coverage estimates are pushed towards an overcount.

The remainder of this paper will discuss the methodology to identify the misclassification of duplicate pairs in the 2020 PES and an adjustment to correct for this bias. As previously mentioned, the 2000 post-enumeration survey estimates were revised in part to correct for a misclassification of duplicate enumerations (U.S. Census Bureau, 2004). The methods I present here are similar to those used in 2000, though not as rigorous because of the time and resource constraints for conducting my analysis.

3. Methodology

I used the United States person E sample and focused on those records that were part of a duplicate pair. This included records where the sample case was determined to be the primary enumeration and those where the sample case was determined to be the duplicate enumeration. For both cases, the data provided the link to the census record for the other half of the duplicate pair. Thus, I could obtain the geographic and demographic information for both halves of the duplicate pair. I focused on those duplicate pairs for which both records were not in the sample block. When both records were in the sample block, the E sample either included both enumerations (and thus had a balance of one primary and one duplicate enumeration), or the PES implemented an adjustment to the correct enumeration probability of the sample record that ensured an unbiased estimate of the correct enumeration rate (Beaghen et al. 2022).

I then removed some other types of duplicate pairs from my analysis. One example is when the sample enumeration had a duplicate link to a record in a group quarters facility, such as a college dormitory or prison. Recall that group quarters are out of scope for the PES, so all PES sample enumerations are for people in a housing unit. The census residence rules generally state that people should be counted in group quarters facilities (U.S. Census Bureau, 2018), so I assumed the group quarters record to be the correct enumeration with certainty. While this impacted a small number of records, it would be worthwhile for future research to investigate why the PES classified the sample enumeration as the primary half of the duplicate pair and whether we should employ a blanket rule to make these sample enumerations be the duplicate in this situation.

Another example of duplicate pairs I removed from consideration were primary (or duplicate) records that were determined to be erroneous because they were out of scope or fictitious. These reasons would apply to both halves of the duplicate pair, and I assumed that the PES would have made the same determination had the duplicate record been selected in the sample. Although a duplicate link existed, these records are erroneous for other reasons, and I did not want such records factored into the adjustment for duplicates.

Similarly, for cases where the original correct enumeration status was unresolved, I removed these cases from consideration and set the correct enumeration probability to 0.5. Here I assumed that the only uncertainty about the case was which half of the duplicate pair was correct, and that there was no uncertainty as to whether the case was erroneous for other reasons, such as being out of scope or fictious.

After these restrictions to my analysis universe, what remained was a set of duplicate pairs for which both enumerations were in a housing unit, those housing units were in different census blocks, and one of the enumerations in each duplicate pair was a correct enumeration with certainty (i.e., a person who should have been counted at that location). Using these cases, I first produced weighted tallies of the duplicate pairs by whether the sample case was the primary or duplicate record. These results showed that the sample case was more often determined to be the primary (correct) enumeration. See Section 4 for the results. I produced these tables by select demographic groups to investigate whether this issue was more pronounced for certain groups.

The next step was to correct for this misclassification and analyze its impact on the estimates of net coverage error. For these duplicate pairs, I calculated an adjustment factor δ to be applied to the primary cases such that after the adjustment, the weighted total of the primary enumerations would be exactly half the weighted total of this set of cases that were part of a duplicate pair. That is, this adjustment factor would force the fifty-fifty split of primary and duplicate weight in the sample.

$$\delta = \frac{0.5 \sum_{i \in D} w_i}{\sum_{i \in D} p_i w_i}$$

Where
D is the set of cases that are part of a duplicate pair for this adjustment, indexed by i,
w_i is the sampling weight
p_i is in indicator for whether the case is a primary (=1) or duplicate (=0).

Since the sample cases were more often classified as the primary enumeration, the adjustment factor was less than 1. To analyze its impact on the DSEs, I multiplied the adjustment factor by the correct enumeration probability for the primary cases, thus decreasing the correct enumeration rate. Since these primary cases were originally correct enumerations with certainty, the new correct enumeration probability was equal to the adjustment factor.

Finally, I used this adjusted E sample to recalculate the DSEs. I ran the same models as the production results, as described in Heim (2022). I did not make any changes to the data-defied model nor the P-sample match model, so those components of the DSE were the same as used in the production results. I produced tables comparing the net error rates for certain demographic domains to the production results. The Limitations section provides some discussion about the decision to not change the P-sample match statuses for cases matching to a duplicated census record. The implication is that the research DSEs presented in this paper may have been lowered too much.

4. Results

Table 1 presents the classification of the duplicate pairs in my analysis universe, as described in the previous section. The table shows the percent of duplicate pairs for which

the sample case was determined to be the primary (correct) enumeration and the percent for which the sample case was determined to be the duplicate (erroneous) enumeration. The total is weighted by the sampling weight. The table provides the results by the geographic distance of the two halves of the duplicate pair (same county, same state but different county, and different state). Note that for a given row, the standard error provided applies to both percent estimates because these percentages sum to 100.

Recall that I removed certain duplicate pairs from consideration, such as those that were within the same block. The 2020 PES estimated 5.2 million duplicates in the census (Khubba et al., 2022), which would imply about 10.4 million records that are part of a duplicate pair. After my restrictions, there were 5.6 million weighted total cases that were part of a duplicate pair for my analysis. Much of the difference is because I excluded duplicate pairs within the same block.

Overall, the sample case was determined to be the primary enumeration for about 63 percent of the duplicate pairs (standard error 0.8 percent). This is larger than the expected 50 percent and suggests a misclassification of the primary and duplicate enumerations for some of the duplicate pairs. Table 1 shows that the percentages are similar across the geographic distances. Appendix Table 1 provides these results by demographic groups and shows that while there is some variation in the rate, each group considered has the result that the sample cases were more often determined to be the primary enumeration.

Table 1. Classification of Duplicate Pairs by Geographic Distance

Geographic distance of the duplicate pair	Weighted total	Percent where sample case is primary	Percent where sample case is duplicate	Standard Error
Total	5,550,000	62.7	37.3	0.8
Within same county	2,692,000	64.6	35.4	1.4
Within same state	1,185,000	60.5	39.5	1.6
Different state	1,673,000	61.1	38.9	1.7

Note: For a given row, the standard error is the same for each percent because these sum to 100 percent.
Source: U.S. Census Bureau, Decennial Statistical Studies Division, 2020 Post-Enumeration Survey

Table 2 provides the information used to calculate the adjustment factor. Of the 5.5 million cases that are part of a duplicate pair in my analysis universe, 3.5 million (the 62.7 percent from Table 1) were determined to be the primary enumeration. The resulting adjustment factor is about 0.8. While this factor could be calculated separately for different subgroups, the results in Appendix Table 1 show that the factor would be relatively constant across those demographic characteristics.

Table 2. Values for Adjustment Factor, δ

Description	Value
Total weighted cases	5,550,000
Half total weighted cases	2,775,000
Weighted primary cases	3,479,000
Adjustment factor, δ	0.7976

Source: U.S. Census Bureau, Decennial Statistical Studies Division, 2020 Post-Enumeration Survey

I used this adjustment factor to decrease the correct enumeration probability of the primary cases so that the set of duplicate pairs in the E sample would be half correct and half erroneous. Using these adjusted probabilities, I reran the dual-system estimation and compared the new research estimates to the production results. Table 3 shows the change in the DSE at the national level. The DSE decreased by about 1.1 million people, resulting in a net coverage error of 368,000 people (standard error 802,000). Like the production result, this research estimate was not statistically significantly different from zero.

Table 3. National Estimates of Net Coverage Error for the United States Household Population

Method	Census	Dual-System Estimate	Net Coverage Error	Standard Error
Production	323,200,000	323,900,000	-782,000	821,000
Research	323,200,000	322,800,000	368,000	802,000

Note: A negative (positive) estimate of net coverage error indicates a net undercount (overcount).
Source: U.S. Census Bureau, Decennial Statistical Studies Division, 2020 Post-Enumeration Survey

Appendix Table 2 provides the results by demographic characteristics. The decrease in the DSEs is relatively uniform across the groups. This was expected, given the results in Appendix Table 1. Thus, we observe the same pattern of coverage error as with the production results. Two groups no longer show statistically significant net undercounts: 18-to-29 females and American Indian and Alaska Natives. In the production results, these groups had estimates that were near the 90 percent confidence threshold, so the decrease in the DSE (with little change in the standard error) gave a new result that no longer met that significance level.

Returning to the group that motivated this analysis, we see that the net coverage error estimates for 18-to-29 males and females continue to show different coverage patterns than Demographic Analysis. Table 4 provides these results. Clearly the issues examined in this paper do not explain all of the difference in the coverage estimates for this age group, and more research needs to be done to understand the differences.

Table 4. Net Coverage Error Rates for 18-to-29 Males and Females (In Percent)

	DSE Production		DSE Research		DA		
Group	Estimate	Standard Error	Estimate	Standard Error	Low	Middle	High
18-to-29 males	*-2.25	0.57	*-1.72	0.56	0.7	0.1	-0.3
18-to-29 females	*-0.98	0.58	-0.44	0.57	1.8	1.3	0.9

* Denotes a percent net coverage error that is significantly different from zero at the 10 percent level using a two-tailed test.
Note: A negative (positive) estimate of net coverage error indicates a net undercount (overcount).
Source: U.S. Census Bureau, 2020 Post-Enumeration Survey and Demographic Analysis

5. Limitations

A limitation of this analysis is that I did not account for any necessary changes to the P sample. While the P sample is independent from the census in terms of its genesis, the treatment of people who are determined to represent the same individual must agree between the two systems. Namely, a P-sample person can only match to a correct enumeration in the census. For example, suppose a P-sample person matched to an E-sample person that was part of a duplicate pair. If the P-sample person was determined

to be a have lived at this location on Census Day, then the record of the duplicate pair in the E sample must be the primary (i.e., correct) enumeration because we are saying that this individual was present in the sample block on Census Day. If the P-sample person was determined to be a mover (i.e., the person lived outside the block on Census Day), then the E-sample record of the duplicate pair cannot be the primary because we are saying that this individual did not live here on Census Day.

In some situations where I have changed the primary and duplicate assignment of duplicate pairs, I don't believe the P sample need be altered. For example, when I used my adjustment factor to decrease the correct enumeration probability of an E-sample case that was determined to be the primary, I am saying that there is some probability that this person actually lived elsewhere on Census Day, namely at the address of the duplicate census record. A P-sample person may have matched to this primary enumeration, which it now less than a 100 percent correct enumeration. However, the match status of this P-sample person is not questioned. Rather, I am now treating this person as having an unresolved mover probability. Because the 2020 PES used procedure B which includes nonmovers and inmovers in the P sample, I think that it is valid to still treat this person as a full match. (Refer to Marra and Kennel (2022) for information about the treatment of movers in the 2020 PES.) However I acknowledge that this treatment is not correct in all situations. Consider a child in a joint custody situation who is counted at the address of each parent. This child did not move between the census and PES reference days, but rather routinely cycles between the two addresses. It may not be appropriate to treat the P-sample record for this child as a mover. Instead, it may make more sense to lower the match probability of the P-sample record to account for the fact that the matching E-sample record is now less than a 100 percent correct enumeration. The lower match probability would increase the DSE, thus counteracting some of the decrease that was seen in my research estimates. Thus, the research DSEs may have been lowered by too much and can be thought of as an extreme adjustment.

A situation that should be investigated further is when the E-sample primary had a duplicate in a group quarters facility. In these situations, I made the E-sample case be the duplicate (i.e., erroneous) enumeration. If there was a P-sample person who matched to this case, then this person should be treated as out-of-scope for the P sample because they were not in the household population on Census Day. I did not make such corrections.

Another limitation of this analysis is my placement of the adjustments at then end of the E-sample processing. I did not redo the imputation for unresolved cases, but rather used the imputations from the production processing. Because my adjustments changed the correct enumeration status of cases, these adjustments would change the donor pool for imputations. In particular, my adjustments removed correct enumerations, which would lower the imputed correct enumeration probabilities for unresolved cases and further decrease the DSEs.

Finally, while the departure from the theoretical fifty-fifty split of primary and duplicate enumerations is useful for identifying that misclassification occurred in the aggregate, we do not have any information to identify which specific records have been misclassified. Ideally, we would like to adjust the data at the person-record level instead of applying the same general adjustment to the whole population. However, given the results in Appendix Table 1 that the rate of misclassification is relatively constant across key demographic groups, I suspect that we would see similar results if we were able to correct the data at the record level.

6. Discussion

This analysis has shown that the 2020 E sample has likely overstated the number of primary enumerations for the set of duplicate pairs. We know that we should see a balance of primary and duplicate enumerations in the E sample. What we don't know is why this error occurred. One possibility is that people in the PES sample may have misreported their Census Day residency. In particular, people may have reported themselves as living in the sample block on Census Day when in truth they were living somewhere else. Marra and Khubba (2022) discuss how this recall bias may have been exacerbated by the delays to the PES schedule because of to the COVID-19 pandemic. If people misreported as living in the sample block on Census Day, then this would have caused us to assign the sample half of the duplicate pair as the primary enumeration.

Whether our result is because of recall bias, measurement error, or errors in the clerical matching is not clear, but we have strong evidence that there was an issue because we should have observed the fifty-fifty split of primary and duplicate enumerations. Further, the evidence of error in the duplicate classification has implications for cases that were not part of a duplicate pair. For people who were counted once, the PES could still have made errors regarding where they should have been counted. That is, I suspect that the underlying issues that led to the misclassification of duplicates pairs, as discussed in this paper, also caused us to misclassify the living situations of other cases in the PES samples. If we misclassified people as having lived in the sample block on Census Day when in fact they lived elsewhere, there are two impacts on the dual-system estimate.

First, for the E sample, we would have overstated correct enumerations by saying that people who were counted in the census in the sample block did in fact live here when they may not have. Since correct enumerations are in the numerator of the DSE, this would cause an overestimate by the DSE. The second impact is for the P sample, the independent sample. If the PES misclassified people as living in the sample area on Census Day when they did not, then these people would have been nonmatches because we would not have found their census record in the sample block. Rather, the census record (if it existed) would have been where the person actually lived at the time. The match rate is in the denominator of the DSE, so an understatement of the match rate would also lead to an overestimate by the DSE.

The issue with these other implications is that, unlike the duplicate pairs for which we have the theoretical fifty-fifty split of primary and duplicate enumerations, in these situations we only have the respondent-provided information on which to rely. Thus, it is uncertain for which cases to make any adjustment and how large that adjustment should be. Research using administrative records as a third system may help identify E- and P-sample cases for which the Census Day residency may have been misreported.

Finally, I'd like to note that I have focused on this one source of error. Like all other surveys, the PES is likely to have multiple sources of error, and these may work in different directions. My research was not meant to fix the PES results, but rather show the impact of this one source of error and think about its implications. While my adjustment lowered the DSEs, there may be other errors in the PES that have a downward bias. Corrections to these errors would raise the DSEs. An example is correlation bias. While the 2020 PES did not exhibit the same patterns of correlation bias as previous decades (Heim, 2022), this does not mean that correlation bias is not present in the estimates.

References

Beaghen, M., R. Turner, M. Jost, and E. Marra, "2020 Post-Enumeration Survey Estimation Methods: Missing Data for Person Estimates," DSSD 2020 Census Post-Enumeration Memorandum Series #2020-J-05, U.S. Census Bureau, 2022.

Heim, K., "2020 Post-Enumeration Survey Estimation Methods: Net Coverage Estimation," DSSD 2020 Census Post-Enumeration Survey Memorandum Series #2020-J-07, U.S. Census Bureau, 2022.

Jensen, E. and T. Kennel, "Who was Undercounted, Overcounted in the 2020 Census?" U.S. Census Bureau, 2022, retrieved from <https://www.census.gov/library/stories/2022/03/who-was-undercounted-overcounted-in-2020-census.html>

Khubba, S., J. Hong, and K. Heim, "2020 Post-Enumeration Survey Estimation Report: Summary of Estimates of Coverage for People in the United States," DSSD 2020 Census Post-Enumeration Survey Memorandum Series #2020-G-01, U.S. Census Bureau, 2022.

Konicki, S., "2010 Census Coverage Measurement Estimation Report: Adjustment for Correlation Bias," DSSD 2010 Census Coverage Measurement Memorandum Series #2010-G-11, U.S. Census Bureau, 2012.

Marra, E., and T. Kennel, "2020 Post-Enumeration Survey: Source and Accuracy Statement for 2020 Post-Enumeration Survey," DSSD 2020 Post-Enumeration Survey Memorandum Series #J-01, U.S. Census Bureau, 2022.

Marra, E., and S. Khubba, "COVID-19 Impacts on the Post-Enumeration Survey," in JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association, 2022.

Shores, R., "A.C.E. Revision II: Adjustment for Correlation Bias," DSSD A.C.E. Revision II Memorandum Series #PP-53, U.S. Census Bureau, 2002.

U.S. Census Bureau, "Accuracy and Coverage Evaluation of Census 2000: Design and Methodology," U.S. Census Bureau, 2004.

U.S. Census Bureau, "2020 Census Residence Criteria and Residence Situations," U.S. Census Bureau, 2018, retrieved from <https://www2.census.gov/programs-surveys/decennial/2020/program-management/memo-series/2020-memo-2018_04-appendix.pdf>

Wolter, K.M., "Some Coverage Error Models for Census Data," *Journal of the American Statistical Association*, 81, 1986, pp. 338-353.

Appendix

Appendix Table 1. Classification of Duplicate Pairs by Demographic Characteristics

Demographic group	Percent where sample case is primary	Percent where sample case is duplicate	Standard Error
Total	62.7	37.3	0.8
Age and sex			
0 to 4	67.3	32.7	3.2
5 to 9	65.5	34.7	2.5
10 to 17	67.4	32.7	1.7
18-to-29 males	56.2	43.8	2.3
18-to-29 females	56.9	43.1	2.2
30-to-49 males	63.8	36.2	2.0
30-to-49 females	66.2	33.8	2.2
50-and-over males	61.6	38.4	1.9
50-and-over females	61.1	39.0	1.9
Tenure			
Owner	63.6	36.4	1.2
Renter	61.1	38.9	1.3
Race alone or in combination with one or more other races			
White	62.3	37.7	1.0
Black or African American	62.8	37.2	2.5
Asian	65.9	34.1	3.4
American Indian or Alaska Native	63.2	36.8	4.3
Native Hawaiian or Other Pacific Islander	72.4	27.6	7.3
Some Other Race	67.5	32.4	2.2
Hispanic or Latino	68.0	32.0	2.1

Note: For a given row, the standard error is the same for each percent because these sum to 100 percent.
Source: U.S. Census Bureau, Decennial Statistical Studies Division, 2020 Post-Enumeration Survey

Appendix Table 2. Net Coverage Error Rates for the Household Population in the United States by Demographic Groups (In Percent)

Demographic group	DSE Production		DSE Research	
	Estimate	Standard Error	Estimate	Standard Error
Age and sex				
0 to 4	*-2.79	0.64	*-2.53	0.64
5 to 9	-0.10	0.56	0.48	0.56
10 to 17	-0.21	0.43	0.43	0.43
18-to-29 males	*-2.25	0.57	*-1.72	0.56
18-to-29 females	*-0.98	0.58	-0.44	0.57
30-to-49 males	*-3.05	0.35	*-2.84	0.34
30-to-49 females	0.10	0.36	0.31	0.36
50-and-over males	*0.55	0.25	*0.82	0.25
50-and-over females	*2.63	0.25	*2.91	0.25
Tenure				
Owner	*0.43	0.24	*0.75	0.24
Renter	*-1.48	0.53	*-1.08	0.53
Race alone or in combination with one or more other races				
White	*0.66	0.21	*1.02	0.21
Black or African American	*-3.30	0.61	*-2.94	0.61
Asian	*2.62	0.77	*2.98	0.75
American Indian or Alaska Native	*-0.91	0.54	-0.53	0.54
Native Hawaiian or Other Pacific Islander	1.28	2.11	1.66	2.11
Some Other Race	*-4.34	0.49	*-3.99	0.48
Hispanic or Latino	*-4.99	0.53	*-4.64	0.53

* Denotes a percent net coverage error that is significantly different from zero at the 10 percent level using a two-tailed test.
Note: A negative (positive) estimate of net coverage error indicates a net undercount (overcount).
Source: U.S. Census Bureau, Decennial Statistical Studies Division, 2020 Post-Enumeration Survey