# RESULTS OF THE CENSUS 2000 PRIMARY SELECTION ALGORITHM

**Stephanie Baumgardner**
**U.S. Census Bureau, 4700 Silver Hill Rd., 2409/2, Washington, District of Columbia, 20233**

**KEY WORDS: Primary Selection, Algorithm, census response**

## I. BACKGROUND

### The Purpose of the Primary Selection Algorithm

There were several ways in which to respond to the 2000 Census. Housing unit addresses that received a short form questionnaire with a 22 digit Census ID had the choice of mailing back the questionnaire or completing the form on the Internet. Respondents had access to Be Counted Forms (BCF) to use if they were concerned that the census had missed them. The Nonresponse Followup (NRFU) field operation collected information from addresses (Census IDs) for which a mailback questionnaire was not received and checked-in by a specified date. The Coverage Improvement Followup (CIFU) field operation focused on Census IDs identified as vacant or delete in the NRFU operation. While these methods, and others, of collecting population data were implemented with the desire of obtaining a more accurate census count, the various methods also presented the possibility of receiving multiple responses from a single Census ID. The Primary Selection Algorithm (PSA) was the computer program designed to analyze these responses and select from among them the records that it deemed most likely to represent the actual census household.

The PSA was not an unduplication process. Unduplication would require that all person records be compared within a single return and across all returns at all Census IDs. The PSA does not compare person records within a single return and is limited to comparing person records on different returns at the same Census ID. The purpose of the PSA was to identify the unique people that were enumerated across all enumerations for one Census ID.

### The History of the Primary Selection Algorithm

The Primary Selection Algorithm was first developed for the 1990 Census in response to the new flow processing design for housing unit enumeration. Questionnaires were microfilmed by automated cameras and the film read by FACT (Film and Automated Camera Technology), which interpreted the "check-box" marks (Optical Mark Recognition) and indicated the presence of write-in entries. Coverage and content algorithms were then performed on the response records - edits that identified questionnaires whose data showed count inconsistencies, households with more than seven members, and forms with too many required items left unanswered. Questionnaires failing these checks were reviewed by clerks. Through contact with the household, count inconsistencies were corrected and missing information was obtained. These questionnaires were then recycled through the data capture system, creating a second response record, a record that included the additional answers and reconciled information. The PSA in 1990 was designed primarily to select the best capture record for each Census ID, and the algorithm assumed that most of the multiple responses were recycles of the same paper form, not records created by different forms for the same Census ID. However, multiple enumerations of the same Census ID also resulted from the overlap of mail check-in operations and Nonresponse Followup (Love, 1990).

During the 1995 Census Test multiple data captures due to an edit was no longer a cause of multiple returns. However, the mailing of a replacement questionnaire to mail non-respondents accounted for many multiple returns. A new PSA was developed for the 1995 Census Test to accommodate the new census design. Because all names were keyed from the returns, computerized person matching between multiple returns at a Census ID was first performed in the 1995 Census Test version of the PSA (Leslie, 1997).

---

*This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.*

The Primary Selection Algorithm for the Dress Rehearsal was an outgrowth of the algorithm developed and implemented for the 1995 Census Test. Its purpose was to determine which response records would define the composition of enumerated housing units by considering the information available from all records captured for each unit. A single response record was either selected from the capture records for a unit or a composite record was created by selecting components from the multiple records available.

PSA processing during the Dress Rehearsal relied on person matching, using a person matching algorithm. For a housing unit where more than one eligible return was received, a pairwise comparison of all eligible returns at a housing unit was conducted. If at least one person on two returns "matched", then these two returns represented the same PSA household. If none of the person records "matched" between the two returns, the housing unit had more than one PSA household. A basic return for each PSA household was selected based on the type and source of the form that created the return. Other returns for the same PSA household were designated as supplemental returns. All persons on all returns were coded with a status indicating whether to keep or ignore (not use) this person. If two or more PSA households existed for one housing unit, a set of criteria was applied to the PSA households at the Census ID to select one PSA household as the primary PSA household. After the primary PSA household was selected for the Census ID, person records in some other PSA households at the Census ID were considered for inclusion in the census household (Love, 1998).

## II. METHODS AND PROCEDURES USED IN CENSUS 2000

### The Primary Selection Algorithm in Census 2000

The Primary Selection Algorithm implemented in Census 2000 was considerably different than the PSA used in the 1990 Census. The chief cause of multiple returns from a Census ID was very different. In 2000, an edit on returns did not create multiple returns in the process of obtaining a more complete response. There were, however, more ways in which a person could respond to the census. While there has always been multiple returns resulting from the overlap of census operations, the magnitude was limited by clerical control of questionnaires received in the District Offices (DOs) in the 1990 Census and prior censuses. This limitation was nonexistent in 2000. For Census 2000, the ability to capture names from returns and perform person matching between multiple returns at a Census ID thereby creating composite households was perhaps the most substantial change from the PSA used in the 1990 Census.

The PSA for Census 2000 resembled the PSA used in the 1998 Dress Rehearsal. To account for changes from the Dress Rehearsal to the Census 2000 enumeration process, some modifications were made to the PSA process. The Coverage Improvement Followup (CIFU) was a new process introduced to enumerate units identified during Nonresponse Followup (NRFU) as a vacant or deleted unit. The CIFU created multiple forms for these units. The blanket replacement form in mailout areas was eliminated which decreased the number of units with multiple mail responses. The redesign of the Be Counted form (BCF) to allow the respondent to identify the return as a household form instead of a return for an individual purported to reduce the magnitude of the within-household error which occurred when persons on BCFs were incorrectly selected as members of the household found to be living at the address to which the BCF was associated (Love, 2000). In addition, results from the Dress Rehearsal evaluations were used to refine the algorithm, especially the hierarchy of selection criteria.

### The Primary Selection Algorithm process in Census 2000

The PSA operated on one Census ID at a time. Major steps are in order as follows:

**Identify ineligible returns** - Ineligible returns were not subject to further PSA processing. One eligible return at a Census ID is the trivial case for the PSA.

**Perform person matching** - Person matching was sometimes performed at Census IDs with two or more returns. Person matching occurred between person records with sufficient information on different returns at the Census ID. Person matching never occurred between person records on the same return or between person records on returns at different Census IDs.

**Form PSA households** - A PSA household is a set of associated persons at one Census ID. The set may contain no persons (a vacant PSA household), or one or more persons. More than one return may contribute to a single PSA household. Returns that do not have any persons in common (determined by person matching) constitute separate PSA

households. Within a Census ID, all returns with a status of vacant are considered to represent one PSA household (one with no persons). One or more PSA households may be formed at a Census ID.

**Select the basic return of each PSA household** - All PSA households, whether comprised of one or multiple returns, have a return that is designated as the basic return. The basic return for a PSA household provides the housing data and the operational variables on the household level. Non-basic returns in the PSA household are called "other" returns for the PSA household. The basic return is selected by sequentially applying a set of criteria to all the returns that make up the PSA household until one return is selected. The criteria are different depending on whether the PSA household is occupied or vacant. In the trivial case where there is exactly one return for the PSA household, that return is declared the basic return. The exact selection criteria are proprietary information and cannot be disclosed.

**Select the primary PSA household** - One or more PSA households could exist for a Census ID. The primary PSA household is the PSA household that is used in further processing with the basic return of the primary PSA household serving as the return level record along with selected person records (on the basic and "other" returns) comprising that PSA household. When a Census ID has just one PSA household, the designation of the primary PSA household is trivial. When more than one PSA household exists, the primary PSA household is selected by sequentially applying criteria to all of the PSA households until only one PSA household is selected. The exact selection criteria are proprietary information and cannot be disclosed.

**Select additional persons to the primary PSA household** - The final step of the PSA is to select certain person records in other PSA households at the Census ID. This step impacted very few Census IDs in Census 2000 and will not be discussed further in this paper.

## III. RESULTS

### Multiple responses to the 2000 Census

Given the number of ways in which a person could decide to respond to Census 2000, there existed considerable concern regarding how many Census IDs would have multiple responses and how well the PSA rules would do in determining the best records to represent the Census ID. The potential impact that the PSA could have on the overall census numbers was sizeable. Table 1 below shows that the number of Census IDs with multiple responses was less than ten percent of all Census IDs.

**Number of returns per Census ID (Table 1)**

| Number of returns | Number | Percent |
|---|---|---|
| 1 | 107,305,027 | 90.54 |
| 2 | 10,740,311 | 9.06 |
| 3 or more | 473,635 | 0.40 |
| **Total** | **118,518,973** | **100.00** |

From the table above it is evident that, of the Census IDs with multiple returns, most were enumerated by two returns. About 55 percent of all Census IDs enumerated by two returns are the result of two enumerator returns. Just over 82 percent of these are the result of one return from NRFU and one return from CIFU and about 15 percent of these result from two returns from NRFU. About a third of all Census IDs with two returns consist of one mail and one enumerator return and about 96 percent of these result from a mailback return (a return physically mailed back by the respondent) combined with a return from NRFU.

### Ineligible returns

The Primary Selection Algorithm defined some returns as ineligible. There are three situations in which a return could be categorized as ineligible. These situations include "blank" returns, other enumerator returns when one enumerator return is marked as a replacement for the others, and enumerator returns identifying Census IDs as deletes determined by enumerators. "Blank" returns (defined as such in processing prior to the PSA) are those that contain no or very minimal information. Enumerator returns designated as replacement returns are intended to replace other enumerator returns resulting from the same operation. Replacement

enumerator returns are generally due to the poor quality of an enumerators' work where that enumerator's workload is redone. For example, two enumerator returns could be completed for a Census ID by the NRFU operation. If one of these returns has the replacement box marked by the enumerator, this return is meant to take the place of the other return and so the other return is ineligible for the PSA. Enumerator returns for Census IDs that an enumerator has determined as not identifying unique census housing units have been classified as deletes (defined as such in processing prior to the PSA) and are not eligible for the PSA.

Table 2 below shows there is a total of 2,656,951 ineligible returns at all Census IDs and the reason they are ineligible.

**Ineligible returns for the PSA (Table 2)**

| Reason for Ineligibility | Number | Percent |
|---|---|---|
| Replaced enumerator return | 696,691 | 26.22 |
| "Blank" return | 176,903 | 6.66 |
| Delete | 1,783,357 | 67.12 |
| **Total** | **2,656,951** | **100.00** |

Since the PSA operates only on eligible returns, the following table shows how many eligible returns were received per Census ID. The following table is exactly the same as Table 1 but shows the number of eligible returns per Census ID instead of the number of all returns per Census ID. Not surprisingly, the number of Census IDs with one return is higher than the same number in Table 1 and the number of Census IDs with more than one return is lower than the same numbers in Table 1.

**Number of eligible returns per Census ID (Table 3)**

| Number of returns | Number | Percent |
|---|---|---|
| 0 | 158,530 | 0.13 |
| 1 | 109,400,198 | 92.31 |
| 2 | 8,716,359 | 7.35 |
| 3 or more | 243,886 | 0.21 |
| **Total** | **118,518,973** | **100.00** |

Of Census IDs with two eligible returns, almost 46 percent result from two enumerator returns and about 88 percent of these are the result of one return from NRFU and one return from CIFU which is expected due to the design of the CIFU operation. About 40 percent of Census IDs with two eligible returns result from one mail return and one enumerator return and about 96 percent of these result from a mail return combined with a NRFU return. These are likely cases of Census IDs returning their mail returns too late (or at least the mail returns being processed too late) and also being enumerated by the NRFU operation.

**Frequency of person matching among returns**

When a Census ID was enumerated by more than one eligible return, person matching was sometimes performed. Person matching occurred between person records on different returns at a Census ID. Person matching never occurred between person records on the same return or between person records on returns at different Census IDs. Only person records that are designated as searchable were qualified for the person matching process. Allowing only searchable person records to be subject to person matching guarded against false matches of person records with little name information and similar characteristics such as sex, race, and hispanic origin. For person matching to be performed at a Census ID, there must exist at least two eligible returns each containing at least one searchable person record at that Census ID. Person matching was performed at about 50 percent of Census IDs enumerated by more than one eligible return. Over 86 percent of Census IDs enumerated by more than one eligible return where no person matching was performed have at least one vacant return.

**Number of PSA households formed**

Table 4 below shows, for all Census IDs, the number of Census IDs with one, two, and three or more PSA household(s). Most Census IDs (over 73 percent) with multiple eligible returns have just one PSA household. Two or more PSA households were formed at just over two percent of all Census IDs. When there are no eligible returns for a Census ID, no PSA household was formed. This occurs in 158,530 (0.13 percent) Census IDs and these Census IDs are not represented in this table.

**Number of PSA households per Census ID (Table 4)**

| Number of PSA households | Number | Percent |
|---|---|---|
| 1 | 115,964,314 | 97.98 |
| 2 | 2,349,988 | 1.98 |
| 3 or more | 46,141 | 0.04 |
| **Total** | **118,360,443** | **100.00** |

**Number of returns forming PSA households**

One or more returns can form a PSA household. The following table shows the number of returns forming PSA households at Census IDs with more than one eligible return. Most PSA households (over 99 percent) at Census IDs with multiple returns consist of one or two returns; over 40 percent are PSA households with one return and close to 60 percent are PSA households with two returns.

**Number of returns forming PSA households at Census IDs with multiple returns (Table 5)**

| Returns forming PSA households | Number of PSA Households | Percent |
|---|---|---|
| 1 | 4,760,140 | 41.66 |
| 2 | 6,561,984 | 57.42 |
| 3 | 97,778 | 0.86 |
| 4 - 9 | 6,953 | 0.06 |
| 10 - 19 | 84 | 0.00 |
| 20 or more | 23 | 0.00 |
| **Total** | **11,426,962** | **100.00** |

**The selection of a vacant PSA household**

The PSA was designed so that a vacant PSA household could be selected as the primary PSA household over an occupied PSA household at the Census ID. While the specifics of how this may happen cannot be discussed, these results are summarized below.

Of Census IDs with two or more PSA households, about 52 percent have a vacant PSA household. Of those Census IDs with two or more PSA households where one is vacant, the vacant PSA household is selected in about 16 percent of cases.

**Comparison of returns received at Census IDs with two returns**

The similarities and differences between person records on different returns at a Census ID can be determined through person matching. When there are two returns at a Census ID and there exists at least one person record between the two returns that match, these returns form a PSA household. Within that PSA household, one return is selected as the basic return. If the other return contains only person records found on the basic return, the other return is defined as a redundant return. A vacant return can also be considered redundant if the basic return of the PSA household is vacant (all vacant returns at a Census ID form one PSA household). When the other return is not redundant, it could still be a part of the PSA household if at least one person record matches between the two returns. When there are no person matches between the two returns, the two returns form two PSA households indicating that the two returns may represent two different sets of persons at the Census ID.

Of Census IDs with two eligible returns in Census 2000, about 70 percent had a redundant return and nearly 57 percent of the redundant returns were occupied. Of Census IDs with two eligible returns and no redundant return, about 86 percent formed two PSA households (no persons in common); two occupied PSA households about 48 percent of the time and one vacant and one occupied household about 52 percent of the time.

**The source of redundant returns**

The manner in which the PSA selected the basic return of a PSA household made it likely that an enumerator return would be classified as a redundant return when it contained the same information as a mail return at the Census ID. Since many of the multiple return Census IDs have an enumerator return involved (combined most often with a mail return or another enumerator return), it is not surprising that almost 85 percent of all redundant returns are enumerator returns. What this suggests is that these enumerator returns are duplicating information already collected for the Census ID and are unnecessary. In fact, more than 55 percent of redundant enumerator returns result from NRFU and nearly 88 percent of these are occupied indicating the likelihood of the receipt of a late mail return. Also, about 43 percent of redundant enumerator returns result from CIFU and

97 percent of these are vacant as expected due to the design of the CIFU operation.

As mentioned above, of all Census IDs with two eligible returns, about 70 percent have a redundant return. The following further describes these Census IDs:

- About 45 percent have two enumerator returns where one is redundant of the other.
  - ➜ Just over 80 percent of these have CIFU and NRFU returns where the CIFU return is redundant.
  - ➜ Another 12 percent of these have CIFU and NRFU returns where the NRFU return is redundant.
  - ➜ About four percent of these have two NRFU returns where one is redundant.

- About 40 percent have a redundant enumerator return and the other return is a mail return.
  - ➜ Over 96 percent of these have a NRFU return and a mailback return where the NRFU return is redundant.

## IV. CONCLUSIONS

In an effort to improve the response rate to Census 2000, operations existed that sometimes provided multiple enumerations of the same Census ID. We found that Census IDs with multiple eligible returns was less than 8 percent in Census 2000 which is lower than what was anticipated. The data also show that we received multiple returns not because respondents chose to respond in a number of different ways but because there were overlapping census operations that were both intentional and unintentional.

In the future, steps should be taken to limit this amount of overlap between census operations. Updating the NRFU universe as late mail returns are received would lower the number of Census IDs with multiple returns and reduce the workload for the NRFU operation, which, in turn, would reduce the cost of this expensive field operation. Research is currently being conducted to determine ways in which new technology could be used to aid in this task. Also, redesigning the CIFU operation so that an additional return for a Census ID is generated only in certain cases would also lower the number of Census IDs with multiple returns.

While the number of Census IDs with multiple returns was low in Census 2000, we should continue to strive for lowering it even more so that a program such as the PSA does not have to make the tough decisions about which records will ultimately represent the Census ID.

## References

Leslie, Theresa, DSSD 1995 Census Test Memorandum Series #4, "1995 Primary Selection Algorithm." United States Department of Commerce, Bureau of the Census. December 1997.

Love, Susan P., "Description of the Decennial Census Algorithm for the Selection of the Primary and Supplemental Records from the 1990 FOSDIC Data Capture Files." United States Department of Commerce, Bureau of the Census. November 1990.

Love, Susan P., DSSD Census 2000 Dress Rehearsal Memorandum Series A-29, Rev. 2, "Revised Specifications for the Within-Block Search and Primary Selection Algorithm in the 1998 Dress Rehearsal." United States Department of Commerce, Bureau of the Census. September 1998.

Love, Susan P., DSSD Census 2000 Procedures and Operations Memorandum Series No. C-2, "Software Requirements Specification for the Census 2000 Primary Selection Algorithm", United States Department of Commerce, Bureau of the Census. June 2000.