Characteristics of People Overcounted in the Census

Sarah K. Heimel

U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233

Abstract

The U.S. Census Bureau conducted the 23rd decennial census of the country's population in 2010. The goal of the decennial census is to count every housing unit and every person in the United States, once and only once. However, sometimes people or housing units get counted more than once. For instance, an apartment might initially be included twice in the census if two different unit designations (Apt A versus Apt 1) were received from two different address sources, but the Census Bureau has successfully implemented procedures that reduce and resolve such housing unit duplication. Duplication of persons is more complex and challenging to resolve than housing unit duplication due to the complexity of living situations as well as the privacy and confidentiality concerns that constrain any attempts to contact and followup with possible duplicates. Person duplication can occur for a variety of reasons; a child with divorced parents might have been counted in the census by each parent, while a person in jail would have been counted in the prison but also could have been counted by their family at home.

In order to address the problem of person duplication, it is essential to understand characteristics of the people who we suspect were duplicated and characteristics of the living quarters where they were counted. Information presented in this paper will include the age of suspected duplicates from the 2010 Census, and comparisons of the two questionnaires where the same person was counted (such as the distance between each address, the number of people duplicated between the two addresses, and whether the same telephone number was provided on each return). By understanding the characteristics of duplicated people, we can continue to research how to prevent duplication on initial census enumerations and how to resolve the duplication that persists.

Key Words:
Census, duplication, overcount, coverage

1.   Background

The U.S. Census Bureau conducted the 23rd decennial census of the country's population in 2010. Counting each person once, only once, and in the right place is the foundation of the decennial census. Oftentimes though, people have multiple places where they spend time and so could be enumerated at more than one place, creating duplication in the census.

People can be duplicated in the census for reasons related either to their living situation (called person-level duplication) or for reasons related to the physical address at which

they live (called housing-level duplication). This paper will focus on person-level duplication.

In person-level duplication, a person may have been included on more than one questionnaire for reasons such as:
- Joint custody situations,
- Enrollment in college,
- Ownership of multiple residences, or,
- Other reasons that led to part-time residency situations.

The Census Bureau has a rule to define the one place where people should be counted:

> *Count people at their usual residence, which is the place where they live and sleep most of the time. People in certain types of group quarters (GQs) on Census Day should be counted at the GQ. People who do not have a usual residence or cannot determine a usual residence, and who are not in one of the certain GQs types, should be counted where they are on Census Day.*

A brief summary of the residence rule with select situations was printed at the top of the census questionnaire that was mailed out, and is shown in Figure 1.

Figure 1: Residence Rule Summary on 2010 Census Mail Questionnaire
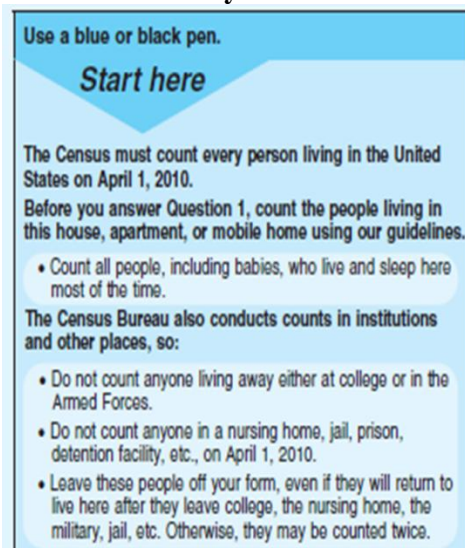
Respondents do not always follow our instructions however. They either do not read them, misunderstand them, or ignore them entirely, resulting in duplication (or other coverage issues). The residence rule was also introduced differently when an enumerator conducted an interview in-person at a housing unit, and there is evidence that the concept was distorted or omitted by some enumerators, also leading to some duplication.

The Census Bureau has developed computer matching algorithms that match the census universe against itself and identify potentially duplicated persons. The algorithms use characteristics such as first name, last name, middle initial, age, date of birth, and phone number to match people. The process involves multiple passes of the system where the matching parameters and constraints are varied for each pass. Each time a person record is matched to another person record, it is given a score that reflects the strength of the

match. The scores are then ranked and the matches are reviewed to establish a cut-off point. Cut-offs are set very high during the review to establish a high level of certainty that only true duplicates and not false matches are identified. All matches with scores above the cut-off are considered to be duplicate person records. The computer matching process identifies duplicates but no individuals are removed from the census during this particular process.

Although extensive research has been done to ensure that chance agreements of name and date of birth are not classified as matches, and while the cut-offs are high, there is still the possibility that persons matched as potential duplicates are not actual duplicates (Fay 2002, Fay 2004, Ikeda and Porter 2007, Ikeda and Porter 2008, Yancey and Winkler 2002, and Yancey 2007). Also, computer matching will fail to identify some duplicates because of inaccurate or missing names or dates of birth. Generally, the Census Bureau prefers to be conservative and not identify some duplicated people rather than identify false matches. This conservative approach allows our limited resources to focus on resolving links that we suspect to be true duplicates.

The computer matching algorithm identifies an association of one person to another, called a "link." The Census Bureau is interested both in the individuals who are linked and in the housing units occupied by those individuals. Two linked person records are considered to be a "person link." The two linked housing units (HUs) involved in a person link are known as "housing unit links." The census questionnaires (or responses) that enumerate the linked people are known as "response links."

Each person link found by the computer matching is classified based on whether the person records were found in housing units or in Group Quarters (GQs). Person links can either be found in two distinct housing units, called HU-HU matches, or between a housing unit and a Group Quarters, called HU-GQ matches. No matching is done to identify person links between two different GQs because the enumeration that takes place at GQs is the final enumeration outcome by design of the operation and there is no mechanism in place to resolve such duplication.

Additionally, all links are classified based on the distance between the housing units. There are five levels at which links can be geographically associated. The levels below are mutually exclusive and are listed in order of precedence:
   a. Within the same block
   b. Within two different but adjacent blocks, called surrounding blocks
   c. Within the same county
   d. Within the same state
   e. Across state lines

Within-block and within-surrounding-block links are considered to be housing-level duplication issues based on our previous research (Pennington 2005) and are thought best resolved in a field operation focused on address resolution[1]. Links beyond the

---

[1] Housing-level duplication might occur when a housing unit is listed twice on the Census Bureau's Master Address File, each listing being slightly different from the other. For instance, an apartment might be identified as Unit A in one listing and as Apt 1 in another listing. The housing unit would have received two Census questionnaires in this case and possibly enumerated themselves on both.

surrounding block are considered to be a result of person-level duplication issues and best resolved with an interview focused on living situations.

2. Limitations

This study of duplicated persons is not the official estimate of duplication in the census since the universe of census returns that was used for this research was not an exhaustive and final universe. Due to timing and other concerns, not all questionnaires were eligible and included in the matching algorithm. For instance, any returns submitted through the Be Counted program or data captured in August were ineligible for the matching.

Categorizing geographic proximity of links (within county, within state, etc) is an imperfect system. For example, a person enumerated at a housing unit in Washington, DC, and at another housing unit in Arlington, VA, is considered a long-distance duplicate because the two responses are located in different states, though the exact addresses could be within a mile of each other.

3. Results[2]

This section describes the universe of suspected duplicated persons, as known at a certain point in census processing. A variety of characteristics will be presented to illustrate the problem of person duplication in the U.S. Census.

Type of Link as either HU-HU or HU-GQ

Table 1 shows the number of response links and the number of person links that were identified, by whether the links were between two housing units or between a housing unit and a GQ.

Table 1: Type of Link

Type of link	Response level		Person level	
	Number	**Percent**	**Number**	**Percent**
HU-HU links	3,857,604	81.9	6,600,215	88.1
HU-GQ links	853,956	18.1	853,956	11.5
Total	**4,711,560**	**100.0**	**7,454,171**	**100.0**

Source: 2010 Census Effectiveness of Unduplication Evaluation

The computer matching algorithm identified 7,454,171 person links from our universe of census returns. The majority (88.1 percent) were identified on two different housing unit responses; the rest were enumerated on one housing unit response and one GQ response. Questionnaires used at GQs only enumerated one person per questionnaire, so by definition there could only be one person from a housing unit response who linked to a distinct GQ response. This explains the constant 853,956 HU-GQ links in each column of Table 1.

The magnitude of person-level links gives a very rough idea of the extent of duplication in the census; we know that some duplication is not identified while some identified links are not actually the same person.

---

[2] All data in this paper is taken from a larger forthcoming Census report, cited in the References section.

Types of GQs

For the 853,956 duplicates that were found in both a housing unit and a GQ, Table 2 shows the type of GQs where they were enumerated.

Table 2: Type of GQ

GQ Type	Number	Percent
College/University Student Housing	441,554	51.7
Nursing Facilities/Skilled-Nursing Facilities	149,306	17.5
Correctional Facilities for Adults	104,351	12.2
Soup Kitchens, Transitional Shelters, Mobile Food Vans	39,807	4.7
Other Non-institutional Facilities[3]	36,744	4.3
Group Homes and Residential Treatment Centers intended for Adults	20,963	2.5
Military Quarters	17,689	2.1
Juvenile Facilities	17,675	2.1
Unknown Group Quarters Type	13,890	1.6
Other Institutional Facilities[4]	11,977	1.4
Total	**853,956**	**100.0**

Source: 2010 Census Effectiveness of Unduplication Evaluation

Over half of the HU-GQ links (51.7 percent) were from college and university student housing. Nursing and skilled nursing facilities accounted for an additional 17.5 percent while correctional facilities for adults accounted for 12.2 percent.

Number of Person Links in HU-HU Links

Table 1 showed that 3,857,604 HU-HU links were identified. Table 3 shows the number of person links that were associated with those HU-HU links.

Table 3: Number of Person Links Within A HU-HU Link

Number of Person Links Per HU-HU Link	Number	Percent
1	2,257,732	58.5
2	909,464	23.6
3	355,574	9.2
4	228,042	5.9
5	101,066	2.6
Over 5	5,726	0.2
Total	**3,857,604**	**100.0**

Source: 2010 Census Effectiveness of Unduplication Evaluation

Over half (58.5 percent) of all HU-HU links had one person duplicated. Almost a quarter (23.6 percent) had two duplicated persons.

---

[3] This category includes living quarters for victims of natural disasters, religious group quarters, U.S. flag maritime/merchant vessels, workers' group living quarters, and Job Corps Centers.
[4] This category includes mental hospitals, hospitals with patients who have no usual home elsewhere, in-patient hospice facilities, military treatment facilities, and residential schools for people with disabilities

Geographic Distance of Link

The Background section discussed the geographic proximity classifications used by the Census Bureau. Table 4 shows how often links were identified in each category.

Table 4: Geographic Proximity of Response Links, By Type

Geography of link	HU-HU Links		HU-GQ Links	
	Number	**Percent**	**Number**	**Percent**
Within block	1,106,807	28.7	93,746	11.0
Surrounding block	360,526	9.3	12,641	1.5
Within county	1,236,974	32.1	243,793	28.5
Within state	699,928	18.1	361,950	42.4
Across state	453,369	11.8	141,826	16.6
Total	**3,857,604**	**100.0**	**853,956**	**100.0**

Source: 2010 Census Effectiveness of Unduplication Evaluation

As mentioned in the background section, within-block links and surrounding-block links are thought to be the result of housing-level duplication issues. These two categories accounted for a combined 38.0 percent of all HU-HU response links. HU-GQ links occurred 11.0 percent of the time within the same block (and 1.5 percent of the time within the surrounding block), which may be an indication that one physical location was listed twice on the Master Address File, once as a HU and then once as a GQ. It can be difficult to determine whether some facilities, such as group homes or nursing facilities, meet the census definition of a GQ or a housing unit.

Within-county links were the most common (32.1 percent) among the HU-HU response links. Within-state links were the most common (42.4 percent) of the HU-GQ response links.

Whole or Partial Household Match

Another way to define response links is by whether they fit the description of being a "whole household" link. Each response link in this evaluation is classified as one of the following:
- Single Person (HH): Both linked returns were from housing units, only one person was enumerated on each return, and that person was identified as a duplicated person. This is essentially a subset of Whole HH matches.
- Whole Household (HH): Both linked returns were from housing units, there was more than one person enumerated on each return, the number of enumerated persons was the same on each return, and all of the persons were identified as duplicates.
- Partial Household (HH): Both linked returns were from housing units, and the number of person links was less than the number of persons on at least one of the returns.
- Discrepant: Both linked returns were from housing units, but the number of person links was greater than the number of persons on at least one of the returns. This could have happened when there were within-return duplicates across both returns. For example, if a single person received two mailback forms (for two slightly different address listings that are really the same place), and put their information down for each of the 12 person panels on each form, then we would

show 12 person links. However, processing would recognize these as within-return duplicates, and thus only one person would be considered valid.

- Single Person HU-GQ: For housing unit to GQ links, the one person enumerated on a GQ return linked to a housing unit with one person enumerated.
- Partial Household HU-GQ: For housing unit to GQ links, the one person enumerated on a GQ return linked to a housing unit with more than one person enumerated.

Table 5 shows the frequency that each of these situations occurred in our universe of census returns.

Table 5: Whole/Partial Status by Type of Link for Response Links

Whole/Partial Description	Number	Percent
HU-HU links	**3,857,604**	**100.0**
Single Person lived and duplicated on each side	412,864	10.7
Whole HH (> 1) lived and duplicated on each side	1,025,139	26.6
Partial HH links	2,417,598	62.7
Discrepant links	2,003	0.1
HU-GQ links	**853,956**	**100.0**
Single Person lived and duplicated on HU side	126,224	14.8
More than one person lived on HU side	727,732	85.2
Total	**4,711,560**	**100.0**

Source: 2010 Census Effectiveness of Unduplication Evaluation

Table 5 shows that 62.7 percent of HU-HU links involved partial links. This indicates that only a subset of the persons at the HUs have complex living situations, such as children in joint custody. For the 37.3 percent of HU-HU links that were either a single person (10.7 percent) or a whole household situation (26.6 percent), resolving the duplication would also mean clarifying the status of the housing unit as potentially vacant on April 1 instead of occupied.

For HU-GQ links, 85.2 percent involved HUs where there were more persons enumerated than just the person at the GQ. Links involving students in college or university housing (shown in Table 2 to be over half of all HU-GQ links) are indicative of this situation.

Phone Number Matches

Every census questionnaire for housing units asked for the respondent's phone number. The increased use of cell phones has been helpful for matching purposes since the phone number listed on linked returns might be the same, even if the respondent was actually at two different addresses. For instance, a household may move to a new address and keep their cell phone as the phone number of record, or a household might have completed a census return for both their primary and secondary home, and listed the same cell phone number on both returns. If a questionnaire was completed by an enumerator using a proxy respondent for an address, then the phone number would have been that of the proxy respondent and not associated with the household members, so those phone numbers were not used in the matching. Thus, it was highly likely that if two forms had matching telephone numbers, then the forms contained the same persons, even if there was some discrepancy in the person data provided on the form.

The following tables indicate how often phone numbers matched between two forms at the response level. Since the GQ forms did not collect a telephone number, the following tables only include HU-HU response links. Blank phone numbers occurred either if a respondent did not provide a phone number, the phone number was found to be invalid, or the respondent was a proxy.

Table 6: Frequency of Phone Numbers Matching (HU-HU links)

Phone Number Agreement	Number	Percent
Both Numbers Matched	949,549	24.6
Phone Numbers Different	2,074,361	53.8
One Number Blank	732,810	19.0
Both Numbers Blank	100,884	2.6
Total	**3,857,604**	**100.0**

Source: 2010 Census Effectiveness of Unduplication Evaluation

Over half of the HU to HU links (53.8 percent) had valid phone numbers on both sides but the numbers were different. About one quarter of the links (24.6 percent) did have the same phone number however. An additional 19.0 percent of the links only had a valid phone number on one side.

Results of Overcount Question

Each person rostered in the census on a housing unit questionnaire could indicate through the overcount question whether they lived or stayed at another location. The overcount question was included after the demographic questions on each person panel. Figure 2 presents the overcount question wording.

Figure 2: Overcount Question from 2010 Census Mail Questionnaire

Since the links in this evaluation represent duplicated persons who seem to have more than one place where they lived or stayed, it would be natural for a high percentage of these links to have positively marked the overcount question on at least one of the responses in a HU-HU link, or on the housing unit side of a HU-GQ link (there was no overcount question on GQ questionnaires). The tables below indicate how often at least one person on a housing unit response indicated an overcount reason; the positive overcount mark might not necessarily have been associated with the duplicated person. Table 7 shows the occurrence of an overcount mark on responses specifically for all HU-HU person links, distributed by geography.

Table 7: Presence of Overcount Mark for HU to HU Person Links, By Geography

Geography	Total HU-HU Person Links		No overcount mark		Had overcount mark	
	Number	Percent	Number	Percent	Number	Percent
Within Block	2,402,030	100.0	2,019,646	84.1	382,384	15.9
Surrounding Block	773,632	100.0	625,955	80.9	147,677	19.1
Within County	1,841,649	100.0	815,551	44.3	1,026,098	55.7
Within State	942,854	100.0	260,598	27.6	682,256	72.4
Across State	640,050	100.0	154,259	24.1	485,791	75.9
Total	**6,600,215**	**100.0**	**3,876,009**	**58.7**	**2,724,206**	**41.3**

Source: 2010 Census Effectiveness of Unduplication Evaluation

Links within the same block or within surrounding blocks did not often include a positive overcount mark; 84.1 percent of within block HU-HU person links had no overcount mark, and 80.9 percent of surrounding block HU-HU person links had no overcount mark. Since these duplications were likely caused by address issues, it was not expected that they would mark the overcount box, since the overcount question pertains to complex living situations and person-level duplication issues.

The overcount question was marked more often in HU-HU links across longer distances; over seventy percent of links across state lines (75.9 percent) and links within the same state (72.4 percent) marked the overcount question.

Table 8 shows the occurrence of an overcount mark on responses for all HU-GQ person links, distributed by geography.

Table 8: Presence of Overcount Mark for HU to GQ Person Links, By Geography

Geography	Total HU-GQ Person Links		No overcount mark		Had overcount mark	
	Number	Percent	Number	Percent	Number	Percent
Within Block	93,746	100.0	80,704	86.1	13,042	13.9
Surrounding Block	12,641	100.0	9,300	73.6	3,341	26.4
Within County	243,793	100.0	147,542	60.5	96,251	39.5
Within State	361,950	100.0	86,537	23.9	275,413	76.1
Across State	141,826	100.0	28,241	19.9	113,585	80.1
Total	**853,956**	**100.0**	**352,324**	**41.3**	**501,632**	**58.7**

Source: 2010 Census Effectiveness of Unduplication Evaluation

The results in Table 8 are similar to those from Table 7. Links between a HU and a GQ within the same block or within surrounding blocks did not often include a positive overcount mark; 86.1 percent of within block HU-GQ person links had no overcount mark, and 73.6 percent of surrounding block HU-HU person links had no overcount mark. The overcount question was marked more often for duplication across longer distances; 76.1 percent of links within state lines and 80.1 percent of links across state lines marked the overcount question. However, 60.5 percent of within county HU-GQ links did not mark the overcount question, a higher rate than observed in HU-HU links for the same geography. This could be a result of the saliency of these living situations; within county HU-GQ links might have originated from short GQ stays, such as a brief period in a local jail or a few weeks in a nursing home for rehabilitation, which were not

permanent enough living situations for respondents to have marked the overcount question.

Age of Duplicated Persons

The demographics of all person links was analyzed, though special consideration had to be taken in comparing information as reported by two different enumerations of the same person. It was possible for the sides of a link to provide different information on the duplicated person or for one side to provide information and the other to not provide any information, even though we considered them to be a match. For instance, one side might have listed a child as being 9 years old and the other side listed the child as being 10 years old. Such cases are labelled as 'inconsistent' in the following table. Additionally, if one side of a link did not provide any response for a certain demographic characteristic, then the person was categorized using the information provided by the other side (assuming it was nonblank).

Age was calculated based on the date of birth provided; if no date of birth was provided, then the write-in age was used when provided. Age was calculated only if the date of birth fell within valid date ranges. Similarly, the calculated age or write-in age was used only if it fell within valid age ranges; otherwise it was considered missing. Missing cases might also represent links where both sides left all age and date of birth boxes blank.

Table 9 shows the distribution of age classifications for all 7,454,171 person links, plus all 308,745,538 people enumerated in the entire census. The data for the final census population has been cleaned so there is no missing data in that column.

Table 9: Reported Age of Duplicated Persons and All Persons in the Census

Age Category	Distribution of Age for Duplicated Persons		Distribution of Age for All Persons in the Census	
	Number	**Percent**	**Number**	**Percent**
Under 5 years	447,959	6.0	20,201,362	6.5
5 to 9 years	526,291	7.1	20,348,657	6.6
10 to 14 years	580,662	7.8	20,677,194	6.7
15 to 19 years	853,916	11.5	22,040,343	7.1
20 to 24 years	878,278	11.8	21,585,999	7.0
25 to 29 years	465,653	6.2	21,101,849	6.8
30 to 34 years	353,585	4.7	19,962,099	6.5
35 to 39 years	325,561	4.4	20,179,642	6.5
40 to 44 years	335,884	4.5	20,890,964	6.8
45 to 49 years	383,374	5.1	22,708,591	7.4
50 to 54 years	399,302	5.4	22,298,125	7.2
55 to 59 years	378,233	5.1	19,664,805	6.4
60 to 64 years	354,145	4.8	16,817,924	5.4
65 to 69 years	280,943	3.8	12,435,263	4.0
70 to 74 years	212,999	2.9	9,278,166	3.0
75 to 79 years	170,811	2.3	7,317,795	2.4
80+ years	323,127	4.3	11,236,760	3.6
Inconsistent	171,474	2.3	NA	NA
Missing	11,974	0.2	NA	NA
Total	**7,454,171**	**100.0**	**308,745,538**	**100.0**

Source: 2010 Census Effectiveness of Unduplication Evaluation

Duplication happens with particular frequency among 15-24 year olds. Of the persons that were identified as potential duplicates, 11.8 percent were 20 to 24 years of age and 11.5 percent were persons 15 to 19 years of age. However, only 7.0 percent of all persons enumerated in the census were 20 to 24 years of age; an additional 7.1 percent were 15 to 19 years of age.

4. Conclusion

Duplication in the census impedes the fundamental goal of counting every person once and only once. Duplication is neither simple to identify nor to resolve, and is not caused by just one or two components of census-taking. This research from the 2010 Census will aid in improving our efforts to identify and resolve duplication in subsequent censuses.

Most of the duplication identified for this research occurred between two housing units, which was expected since the majority of people in the country live in housing units instead of GQs. When duplication did occur to a GQ, college housing accounted for the majority of it. Youth, especially college-aged persons, have disproportionately high rates of duplication when compared to the entire population. Roughly one-quarter of all links between two housing units provided the same non-blank telephone number on the census return, which could be utilized in the future to identify duplication as the enumeration is happening. The overcount question could also be utilized more in the future to identify and resolve duplication, as a majority of links that were found within the same county,

within the same state, or across state lines had positively flagged the overcount question in 2010.

Acknowledgements

I would like to acknowledge Ryan King of the U.S. Census Bureau for helping to analyze this data and Mandy Sha of RTI International for putting this panel together.

References

Fay, R.E. (2002). "Probabilistic Models for Detecting Census Person Duplication," 2002 Proceedings of the Joint Statistical Meetings on CD-ROM, American Statistical Association, Alexandria, VA, pp. 969-974.

Fay, R.E. (2004). "An Analysis of Person Duplication in Census 2000." 2004 Proceedings of the Joint Statistical Meetings on CD_ROM, American Statistical Association, Alexandria, VA, pp. 3478-3485.

Heimel, Sarah (2010). "2010 Census Study Plan for the Effectiveness of Unduplication Evaluation." DSSD 2010 Decennial Census Memorandum Series #O-A-29, U.S. Census Bureau, December 3, 2010.

Heimel, Sarah and King, Ryan (2012), "2010 Census Effectiveness of Unduplication Evaluation Report," To be released.

Ikeda, Michael and Porter, Edward (2007). "Initial Results from a Nationwide BigMatch Matching of 2000 Census Data." SRD Statistics Research Report Series #2007-22, U.S. Census Bureau, December 29, 2007.

Ikeda, Michael and Porter, Edward (2008). "Additional Results from a Nationwide Matching of 2000 Census Data." SRD Statistics Research Report Series #2008-2, U.S. Census Bureau, March 5, 2008.

Lamas, Enrique (2009), "Residence Rule and Residence Situations for the 2010 Census." 2010 Census Information Memoranda Series No. 35, U.S. Census Bureau, April 24, 2009.

Marshall, Leah (2008). "Potential Duplicates in the Census – Methodology and Selection of Cases for Followup." Paper presented at the American Association for Public Opinion Research Annual Conference, New Orleans, LA, May 15, 2008.

Pennington, Robin (2005). "Evaluation of Person Duplication." 2004 Census Test Evaluation 8. U.S. Census Bureau, September 28, 2005.

Yancey, William E. and Winkler, William E. (2002). "Record Linkage Software: User Documentation." U.S. Census Bureau, October 10, 2002.

Yancey, William (2007), "BigMatch: A Program for Extracting Probable Matches from a Large File," SRD Research Report Series #2007-01, U.S. Census Bureau, June 15, 2007.