"Editing Unmarried Couples in Census Bureau Data"

SUMMARY

In Census Bureau classifications, married-couple households only consist of opposite-sex couples, while unmarried partner households may consist of either opposite- or same-sex couples. This classification relies not only on the accuracy of the responses to the household relationship item—either as a spouse or unmarried partner—but also to those on gender. Although gender is usually the most accurately reported item on a survey, minor errors in gender could have a substantial impact on the estimates of same-sex unmarried partner households. This paper outlines some of the issues in estimating the number of same-sex unmarried partner households and the potential effects of these errors on the estimated population.

"Editing Unmarried Couples in Census Bureau Data"

By Martin O'Connell and Gretchen Gooding [1]

Housing and Household Economic Statistics Division, U.S. Census Bureau

Historical Background

One of the most widely discussed household and family tabulations from Census 2000 concerned that of unmarried partner households.[2] Of the 5.5 million unmarried partner households in 2000, 4.9 million were opposite-sex partners while another 0.6 million were same-sex partners. When added to the 54.5 million married-couple households (consisting only of a householder and spouse of the opposite sex), there were a total of 60 million households containing married or unmarried couples.[3]

Crucial to the classification of households into one of these three groups is the joint combination of responses to two items on the form:

1) The relationship of the person to the householder (a spouse or an unmarried partner).

2) The gender of the two people.

---

[1] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on statistical and methodological issues are those of the authors and not necessarily those of the U.S. Census Bureau.

[2] These data generated many reports including a series of analytical papers published by the Urban Institute and a budgetary impact analysis report prepared for the Committee on the Judiciary, U.S. House of Representatives, by the Congressional Budget Office (*The Potential Budgetary Impact of Recognizing Same-Sex Marriages,* June 21, 2004).

[3] Tavia Simmons and Martin O'Connell, *Married Couple and Unmarried Partner Households: 2000,* Census Special Reports, CENSR-5 (US Census Bureau: Washington, DC, 2003).

Although gender in Census 2000 had both the lowest allocation rate (0.9 percent) and index of inconsistency (1.7 percent) of all items on both the short and long forms[4], an analysis of the names of the people may occasionally reveal that their responses are at odds with their reports on gender. Because the number of unmarried-partner households is relatively small, minor errors in gender could have a substantial impact on these estimates. This paper will explore the possible effects of errors in the reporting of the sex item on the size of the unmarried and married-couple populations. The results presented are hypothetical exercises and do not represent revisions to any previously published Census Bureau data.

Decennial Census Editing Procedures

The editing specifications used for Census 2000 stated that if a household consisted of a married couple with both spouses reporting the same sex—and where no imputations were made for either person for either their relationship or sex due to non-response—the partner who reported being a "spouse" of the householder was changed to being an "unmarried partner" of the householder. This was a different process than that used in the 1990 Census where the *relationship* category would have remained the same (spouse), but the *sex* of the partner would have usually been changed.

---

[4] The index of inconsistency is a measure of response variance in questions. Part of the Census 2000 program was to conduct a Content Reinterview Survey to measure the consistency of responses between questions on Census 2000 and a subsequently administered survey. For a description of this survey and the ensuing analysis, see Paula J. Schneider, *Content and Data Quality in Census 2000,* Census 2000 Testing, Experimentation, and Evaluation Program Topic Report No. 12, TR-12 (US Census Bureau: Washington DC, 2004), Table 1.

This change in the editing process for Census 2000 was made for several reasons. As previously noted, individual reports of sex are usually the best reported items on surveys—names could have errors of legibility when being scanned by optical readers and may not be as reliable as a simple single mark on the sex item. In addition, the passage in 1996 of the Federal Defense of Marriage Act (H.R. 3396), included a provision that Federal agencies recognize only persons of the opposite-sex in defining a married couple for Federal program purposes. While the Act did not specify how marital status information should be collected by the Federal government, it did define, for purposes of federal law, "marriage" as a legal union between a man and a women and "spouse" as a person of the opposite sex who is a husband or a wife. However, the edit attempted to preserve the apparent intent of the relationship by assigning the spouse to the unmarried partner category instead of randomly allocating relationship codes based on sex and age. (It should be noted that in the overall editing process of short form items, same-sex partners could also be allocated if responses to the relationship item were left blank on the form.)

The editing process in the Census is very complex. It is an iterative process that compares responses among all household members to ensure that the resulting household does not contain any anomalies (for example, a householder with multiple spouses or children who are older than their parents). Because of this process, an examination of the imputation flags of respondents on data files may provide one with a general picture of how the final imputed value was obtained but will not provide one with a trace record of

all of the possible changes that were made during the process. Public use data files do

not include the original answers given by respondents, only the final "edited" values.

Because the transference of a "same-sex spouse" to a "same-sex unmarried partner" was

accomplished through this assignment process, this type of change to the data was

recorded as a household consistency assignment. This change was not tabulated as an

allocated value or used in any of the imputation rates published in Census 2000 tables.

The allocation flag for the relationship item on the public use file did not indicate that this

type of assignment had been made—it was recorded as "Not allocated." In fact, it was a

general rule that Census 2000 allocation flag variables would not contain any detailed

information on the type of data assignments or allocations made during the editing

procedure, but only if a value had been allocated in its final state.

Issues Faced by Researchers

The decision to release the allocation variable in this restricted format makes it

impossible for researchers to ascertain how many of the originally marked same-sex

spouses were "transferred" in the final editing steps to same-sex unmarried partners.

There have been attempts by researchers to develop proxies for this flag by assuming that

all same-sex unmarried partners that have allocation flags indicating that their marital

status was changed in the editing process were originally recorded as same-sex married

spouses. Using this proxy analysis, some suggest that 30 to 40 percent of all same-sex

unmarried partners are misclassified and *assume that ALL* of these couples are truly

opposite-sex married couples, an assumption that cannot be established given external

researchers access to only public use data files.[5]

 In addition, since the marital status variable was only on the long form, and was not

involved in the determination of couple's status in the short form edit, this process uses

yet another allocation flag for this transference. Marital status allocation flags may be set

in an editing procedure for numerous reasons (for example, blank responses to the item)

other than changes in relationship status. Thus, this indirect analysis is based on several

weak assumptions and at best provides a suggestive yet inconclusive analysis.

However, even if a marker were included on the public use file which indicated that an

assignment had been made, it would fail to answer several key questions:

1. How many of these reassignments were based on incorrect marks to the sex
   question and how many were made because same-sex partners considered
   themselves to be in a spousal living arrangement (that is to say, no error was
   made)?

2. How many same-sex couples, after the edit, are incorrectly recorded as *opposite-sex* married couples or unmarried partners because they too erred in marking the
   sex item (one partner was incorrectly marked as being of the opposite sex)?

3. What are the characteristics of all population groups, both those with and without
   errors or reassignments?

---

[5] Dan Black, Gary Gates, Seth Sanders, and Lowell Taylor, "Same-Sex Unmarried Partner Couples in Census 2000: How Many are Gay and Lesbian?" Paper presented at the conference "Measurement Issues in Family Demography," Bethesda, Md., November 2003.

As a basis of comparison, it should be noted that estimates of same-sex unmarried partners from Census 2000 SF3 (the 2000 Census sample table PCT1) were very close to those provided by the American Community Survey (ACS) from the Census 2000 Supplementary Sample table PCT008 (about 660,000 for each), even though the ACS used a completely different editing and processing system. More importantly, ACS data collected in interviews used telephone and computer-assisted instruments that had verification steps to correct any errors in the reporting of the relationship and sex items. If a person was reported as being the spouse of the householder and the same sex as the householder, a question was asked to either confirm or correct the responses, thereby eliminating any errors in the relationship and sex items during the actual interview even before the processing occurs.[6] The fact that the Census 2000 numbers-- without this sex verification check--came so close to the ACS numbers--that did have this verification procedure--indicates that it would be incorrect to assume that all spouses reassigned to unmarried partners in Census 2000 were the result of errors in marking the sex item on the questionnaire. If that were the case, the ACS numbers would have fallen considerably below the Census 2000 data because of the verification procedure used in the ACS to catch these mistakes.

Framework and Data Requirements

Aside from conducting a prohibitively expensive and time consuming re-interview of every household in the United States to verify both sex and relationship responses, the only economically and statistically feasible way to estimate the number of misclassified

---

[6] CATI/CAPI interviews comprise approximately 40-50 percent of all ACS interviews (weighted cases).

same-sex unmarried partners would be to use a data set containing the first names of the

respondents and the probability that a person's name is associated with a specific gender.

Some of the questions that could be answered in such an analysis are outlined below:

1) Of those partners assigned by the current editing scheme, how many would revert

   to being married opposite-sex spouses and how many would still remain as

   "same-sex spouses" on the basis of their names? It would be incorrect to believe

   that all of the transferred spouses were done so in error and were of the opposite

   sex. There will be same-sex couples that will report themselves as being married

   either because they have gone through a marriage or domestic partner ceremony

   or consider themselves as living together as a married-couple family, especially if

   any children are present. Clearly, there are currently people of the same sex who

   have married in Massachusetts and in countries other than the United States.

2) What are the characteristics of same-sex couples by their transference status? Do

   couples that originally reported themselves as same-sex spouses have different

   demographic and economic characteristics than same-sex couples that originally

   reported themselves as unmarried partners? Are these characteristics indicative of

   differences in family living arrangements such as their age, the presence of

   children in the household or differences in employment?

3) How many same-sex spouses/couples are currently being incorrectly tabulated as

   opposite-sex couples/partners because they too had an error in the marking of the

   sex item on the form? If opposite-sex couples can make errors when marking

their sex and appear to be of the same sex, then same-sex couples can also make

errors and inadvertently be classified as being of the opposite sex.

By using a data file containing the first names of the respondents, one can better examine

the possible gains or losses to different types of coupled households if respondents'

names were used to verify the report of their sex. Of course, one cannot reasonably

expect coding staff to examine every first name of every person in the Census—that

would require examining millions of names of couples and then, at best, coming to a very

subjective decision concerning the likelihood of a name being male or female. What is

the gender of a person named Pat, Leslie, Sean, Jean, Ryan, etc.? Would coders

reviewing names know the gender of people with names of non-European extraction?

The Census Bureau has developed a statistical "name directories," which are files of first

names that are associated with a probability index that identifies the "maleness" of the

name. These name directories were developed for each state from the Census 2000 data

files. The probability index (from 0 to 1000) for each name in the directory was

constructed by taking the ratio of the number of times this name was recorded by a male

to the total number of times this name was recorded by either a male or female.

For example, an index of 950 indicates that when this name appeared in the Census 2000

for a given state, 950 times out of 1000, that person was a man. An index of 20 would

indicate that only 20 times out of 1000 that name was reported by a man or conversely,

980 times out of 1000 that name was identified as being reported by a woman. A

decision, then, could be made as to whether to accept the respondent's reply of their sex on the basis of consistent reports with this index or to reject their response and assign them to the opposite sex. Clearly, age, cultural and geographical differences may affect this probability, as similarly spelled names may be male or female in different cultural environments. Directories prepared at the State level can partly address these issues.

By setting different "acceptance levels" for this index, one can see the effect of using an alternative piece of information—a person's name—in the review or editing of data files.[7] For example, suppose one was very confident that an error was made in marking the sex item as "female" if a person's name 99 percent of the time was recorded as "male" in the names directory. One could reassign sex from female to male for all people who's a name had an index value of 990 times out of 1000 (99 percent).

One could lower the confidence or acceptance level to 950 or 900 times out of 1000, but that would risk making more false assignments. A name more likely to have both male and female responses (for example, Leslie compared with Elizabeth) would have a lower index level. A decision to alter the sex response for names with lower index values would have a greater potential for making errors when assigning people with those names to the opposite sex.

---

[7] Some Census 2000 editing routines did use a person's name to assign a male/female value for the gender item when that question was left blank on the form and no other useful information was available for editing procedures.

Using this type of index[8], one could examine how many same-sex unmarried partner households have partner names that could imply an inconsistency with their gender—that they are likely to be opposite-sex married couples—and hence, the editing procedure may have produced an *overestimate* of same-sex partners. But this analysis also addresses the following issue: How many currently accepted opposite-sex couples (married or unmarried), when using the same verification procedure, would have one partner's sex altered, thus *adding* to the count of same-sex unmarried partners? This type of transition analysis would provide a better measurement of the number of same-sex couples in the United States and clearly answers questions that a simple allocation flag cannot address in any comprehensive fashion. In fact, only having an allocation flag would provide a biased and incomplete analysis of this problem as will be shown in the model below.

Model Estimates of Couple Transitions

The magnitude of revisions to any initial estimate of same-sex couples produced by Census editing routines depends on three components:

1) The size of subpopulations making up the total same-sex population (SST): those being assigned from married spouses (SA) and those not assigned but reporting themselves as same-sex unmarried partners (SN).

2) The size of subpopulations which may still contain same-sex couples but were not identified as such in the edit because they incorrectly marked one partner as being of the opposite sex: this group consists of opposite-sex married couples (MC) and opposite-sex unmarried partners (OS).

---

[8] This index file is not available to the public.

3) The transfer rates—the percentage of couples where one or more partners marked their sex "incorrectly" as determined by a first name analysis. Transfer rates from the assigned (T_{SA}) and not assigned (T_{SN}) same-sex populations would generate *population losses* from these same-sex groups to opposite-sex spouses and opposite-sex partners, respectively. Transfer rates for married couples (T_{MC}) and opposite-sex partners (T_{OS}) would generate *population gains* to same-sex unmarried partners from the two opposite-sex groups.

The revised count of same-sex unmarried partners (SSR) can be estimated using the following model:

$$(1) \quad SSR = SST - (SA*T_{SA} + SN*T_{SN}) + (MC*T_{MC} + OS*T_{OS})$$

In the absence of having a readily accessible data file from Census 2000 with the gender probability index values for first names attached to each record, we can model a range of estimates of the number of same-sex unmarried partners using different scenarios of population sizes and transfer rates from previously published Census data and research papers. The purpose of this exercise is not to produce a new estimate of same-sex unmarried partners but to examine possible ranges of estimates and the sensitivity of population counts to the parameters expressed in the model above.

Population parameters

The base population counts of same-sex unmarried partners (SST), married couples (MC) and opposite-sex unmarried partners (OS) are readily available on the American Factfinder from Census 2000, Summary Tape File 1, tables P18 and PCT14. These data, from the 100 percent short form, are shown in Table 1: there were 594,391 same-sex couples, 54,493,232 married couples, and 4,881,377 opposite-sex couples in Census 2000. Data are not presented in any published Census report on the number of those 594,391 same-sex unmarried partners who were assigned that status because they reported themselves on the form as being of the same-sex and as spouses. However, one can use for this exercise the indirect estimates suggested by Black *et al*. (2003) that 40 percent of those couples were assigned from the initial population of married couples. This produces an estimate of 237,756 assigned couples (SA) and 356,635 not assigned couples (SN).

Transfer rates

Ranges for rates of misreporting of gender for specific types of couples can only be suggested from the Census 2000 Content Reinterview Survey.[9] Data from the content reinterview test indicate that the index of inconsistency for reports of sex was 1.7 percent, lowest of any item on Census 2000. The 2004 test census of New York, which generally covered the borough of Queens,[10] was also used to estimate transfer rates. Results suggested that, using the first name index to evaluate reports of sex at the 99 percent, 95

---

[9] See Schneider, op cit., Table 1.

[10] Martin O'Connell and Gretchen Gooding, "The Use of First Names to Evaluate Reports of Gender and Its Effect on the Distribution of Married and Unmarried Couple Households." Paper presented at the Annual Meetings of the Population Association of America, Los Angeles, CA, March 30-April 1, 2007.

percent, and 90 percent level of acceptability, a range from about 1 to 2 percent of both married couples and opposite-sex unmarried partners are likely to have made a mistake when marking the sex item on the census form that would result in a reassignment of their sex (transfer rates T_{MC} and T_{OS}, respectively).

Using these estimates, we can use as a range of possible transfer rates for opposite-sex couples from 1 percent to 2 percent as shown in Table 1.   This would produce overall gains to the same-sex population from 0.6 million to 1.2 million couples on the basis of using first names to edit the sex item (row 6).

Data from the New York test indicated rates of discrepancy between first names and index levels for those same-sex partners who were not assigned their status in the range of 4 percent to 6 percent (T_{SN}). For same-sex unmarried partners who were assigned from the original pool of married couples, considerably higher transfer rates were chosen for these assigned couples (T_{SA}) ranging from a low of 40 percent to a high of 50 percent, again based on the New York data.

Table 1 presents the model using ranges of transfer rates from the lowest to the highest levels as proposed by previous research. The "Low" and "High" models do not necessarily represent the lowest and highest resulting *numbers* of same-sex couples generated by the model but the lowest and highest levels of *transfers* to the opposite sex when using first names to edit the sex item on the questionnaire.

The resulting model-based estimates shown in Table 1 indicate *in all hypothetical examples*, if an attempt was made to re-distribute the data based on the changes to the sex item using the respondent's first name, the final number of same-sex unmarried partners (SSR) would range from 1.1 million to 1.6 million partners (row 7), compared with the original count of 594,391 partners (row 1). Because of the overwhelming size of the opposite-sex couple population (59 million), even small proportions of sex reassignments, as determined by the use of first names, would produce large additions to the same-sex partner population.

The last column in Table 1 shows the net effects of using different combinations of transfer rates designed to maximize losses to same-sex couples and minimize gains from opposite-sex couples. Under this scenario, the number of same-sex partners generated from name/sex transfers among opposite-sex couples (593,746—row 6) is more than four times the total loss from the same-sex partner categories (140,276—row 3).

Summary

Current Census Bureau editing procedures assign couples of the same-sex that indicate that they are spouses to the category of unmarried partners. This paper has attempted to provide a framework to analyze the potential effects of errors of marking the sex item in questionnaires on the number of same-sex unmarried partners. Recognizing that it would be economically and practically impossible to re-interview every couple in the United States to verify their sex, a model is developed to evaluate the net additions or losses to

the different coupled universes under different levels of confidence when using names to edit the respondent's sex.

Hypothetical examples were developed for varying levels of sex reassignments among the different population groups based on prior analysis of Census test data. In all cases, the net effect of attempting to use first names to verify and subsequently alter the response to the sex items could potentially increase the number of same-sex unmarried partner from the current level of 0.6 million in Census 2000 to a range of 1.1 million to 1.6 million, depending on the assumption. Only if an actual file of Census 2000 households with an associated names probability index was available could this issue be more fully investigated. In addition, that data file would also permit an evaluation of the demographic characteristics of the different populations before and after any revisions were made because of using names to edit the sex item.

References

Dan Black, Gary Gates, Seth Sanders, and Lowell Taylor, "Same-Sex Unmarried Partner Couples in Census 2000: How Many are Gay and Lesbian ?" Paper presented at the conference "Measurement Issues in Family Demography," Bethesda, Md., November, 2003.

Martin O'Connell and Gretchen Gooding, "The Use of First Names to Evaluate Reports of Gender and Its Effect on the Distribution of Married and Unmarried Couple Households." Paper presented at the Annual Meetings of the Population Association of America, Los Angeles, CA, March 30-April 1, 2007.

Paula J. Schneider, *Content and Data Quality in Census 2000,* Census 2000 Testing, Experimentation, and Evaluation Program Topic Report No. 12, TR-12 (US Census Bureau: Washington DC, 2004).

Tavia Simmons and Martin O'Connell, *Married Couple and Unmarried Partner Households: 2000,* Census Special Reports, CENSR-5 (US Census Bureau: Washington, DC, 2003).