Overview of error model for estimates of foreign-born immigration using citizenship and residence one year ago from the American Community Survey

Mary H. Mulry[1]
U.S. Census Bureau

1.Introduction

Demographic Analysis (DA) estimates of the U.S. population on April 1, 2010 included estimates of foreign-born immigration each year between 2000 and 2010 based on data from the American Community Survey (ACS). Our goal is to design methodology to describe the uncertainty in the estimates of foreign-born immigration. The main estimation method we are considering in this paper uses the responses to two questions, one that asks citizenship and another that asks residence one year ago (ROYA). We use an error model that accounts for sampling and nonsampling errors. Unfortunately, time and resource limitations prevent us from conducting studies to measure the nonsampling errors so we intend to propose reasonable estimates based on studies of nonsampling errors in ACS for other purposes or studies of nonsampling errors in other surveys.

Also, the ROYA question in the Puerto Rico Community Survey (PRCS), which is the ACS adapted for Puerto Rico, was used to estimate migration from the U.S. to Puerto Rico starting with the year 2005. Since the PRCS began in 2005, previous research was used to estimate net migration between the U.S. and Puerto Rico for the years 2000 to 2004 (Christenson 2002). Although this paper focuses on the error model for describing the uncertainty in the estimates of foreign-born immigration using the ACS, the same basic principles apply when developing a methodology for the uncertainty in the PRCS estimates of the two components of migration between the U.S. and Puerto Rico.

This paper describes our strategy to use the error model in the design of a simulation to study the propagation of errors in the estimates of foreign-born immigration. The results of the simulation study will produce a sensitivity analysis that assesses the uncertainty in estimates of foreign-born immigration. An important guide in constructing an error model is to identify the assumptions in the estimation and examine where failures in these assumptions potentially occur in data collection, data processing, imputation, and estimation.

2.Form of Estimator for Foreign-born Immigration

The ACS estimation of foreign-born immigration each year includes a ratio adjustment to population control totals from the Population Estimates Program (PEP). The final steps in the calculation of the ACS weights include the formation of 156 cells defined by race/Hispanic ethnicity, sex, and age. The ACS estimates in these cells are controlled to the estimates of the population totals as of July 1 that the PEP produces each year (U.S. Census Bureau 2009).

[1] Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are the authors' and not necessarily those of the U.S. Census Bureau.

DA estimates foreign-born immigration as of July 1 for a year by using all foreign-born ACS people who are at least one year old and who lived outside the U.S. or Puerto Rico one year ago, the year in the survey months January through December. Appendix 1 shows the questions used to identify the foreign-born and those whose residence one year ago was outside the U.S. The foreign-born are determined using Question 8 regarding citizenship and Question 15a regarding residence one year ago. The estimates include an adjustment that accounts for babies less than one year of age by adding half of the estimated number of people with a residence one year ago abroad and a reported age of one year. In addition, the estimates for years 2000 through 2004 are three-year averages since the sample sizes were smaller before the full-scale implementation of ACS began in 2005. Table 1 shows the estimates of foreign-born immigration for 2000 to 2008 including how the three-year averages were implemented.

Table 1. Estimates of foreign-born immigration by year (thousands).

Period: July 1 to June 30 (ACS data year)	Foreign-Born population in the ACS whose residence one year ago	Additional assumption for those under age 1	Total	Foreign-Born Immigration (w/ smoothing for 2000-2004)
(A)	(B)	(C)	(D) = (B) + (C)	(E)
1999-2000 (2000)	1,420	12	1,431	1,431
2000-2001 (2001)	1,421	10	1,431	1,431
2001-2002 (2002)	1,228	13	1,241	1,368
2002-2003 (2003)	1,025	8	1,032	1,235
2003-2004 (2004)	1,124	9	1,134	1,136
2004-2005 (2005)	1,188	8	1,196	1,196
2005-2006 (2006)	1,190	8	1,198	1,198
2006-2007 (2007)	1,114	8	1,122	1,122
2007-2008 (2008)	1,082	8	1,090	1,090
2008-2009 (2008)	1,082	8	1,090	1,090

Source: U.S. Census Bureau (2010)

Next, let us consider the form of the estimator for foreign-born immigration. Define the following for a cell C used in the ratio adjustment:

P = size of population in cell C
F = size of the foreign-born population in cell C
Y = size of foreign-born population in cell C who resided outside the U.S. one year ago
$s_F = F/P$ = proportion of the population within cell C who are foreign-born
$r_Y = Y/F$ = proportion of the foreign-born within cell C who resided outside the U.S. one year ago
$I_F(i) = 1$ if person i is foreign-born, 0 if person i is native-born

$I_Y(i) = 1$ if person i resided outside the U.S. one year ago, 0 if person i lived in the U.S. one year ago

w_i = ACS survey weight for person i (before ratio adjustment).

Then we can use the ACS to form a weighted estimate of the population size, say $\hat{Q}$ as well as estimates of F and Y as follows:

$$\hat{Q} = \sum_{i \in C} w_i$$

$$\hat{F} = \sum_{i \in C} w_i I_F(i)$$

$$\hat{Y} = \sum_{i \in C} w_i I_F(i) I_Y(i) \quad .$$

If we let $\hat{P}$ be the estimate of the population size P for cell C from the Population Estimates Program (PEP) used in the ratio adjustment for cell C, the estimator used by DA for estimating the foreign-born immigration for cell C, $\hat{T}$, has the form

$$\hat{T} = \frac{\hat{P}}{\sum_{i \in C} w_i} \sum_{i \in C} w_i I_F(i) I_Y(i)$$

$$= \frac{\hat{P}}{\sum_{i \in C} w_i} \times \frac{\sum_{i \in C} w_i I_F(i)}{\sum_{i \in C} w_i I_F(i)} \times \sum_{i \in C} w_i I_F(i) I_Y(i)$$

$$= \hat{P} \times \frac{\sum_{i \in C} w_i I_F(i)}{\sum_{i \in C} w_i} \times \frac{\sum_{i \in C} w_i I_F(i) I_Y(i)}{\sum_{i \in C} w_i I_F(i)}$$

$$= \frac{\hat{P} \hat{F} \hat{Y}}{\hat{Q} \hat{F}}$$

$$= \hat{P} \hat{s}_F \hat{r}_Y$$

where $\hat{s}_F$ and $\hat{r}_Y$ are estimators of s_F and r_Y from the ACS using survey weights before the ratio adjustment.

3. Error model for foreign-born immigration

The estimate of the foreign-born immigration may be affected by

- errors in PEP estimates
- ACS data errors that create error in the ACS estimate of the foreign-born who resided outside the U.S. one year ago

- errors caused by an inconsistency between the ACS and the PEP in the variables used to form cells for the ratio adjustment.

Each year, the immigration component of the PEP estimates uses ACS data collected the previous year. Therefore, since the full implementation of the ACS, there has been no overlap in the ACS data used in estimating the foreign-born immigration for two consecutive years.
The error in the estimate of the foreign-born immigration $\hat{T}$ may be expressed in terms of a bias component β and a random error component ε . Then the form of the model is

$\hat{T} = T + \beta + \varepsilon$
where $E(\varepsilon) - 0$.

The bias β may be expressed as the sum of the bias due to inconsistency in the reporting of characteristics between PEP and ACS β_I, and the bias due to error in the data used to form the PEP estimation and the ACS estimation β_D,

$$\beta = \beta_I + \beta_D .$$

The random error ε has terms due to sampling variance ε_S and variance due to the imputation for missing data ε_M,

$$\varepsilon = \varepsilon_S + \varepsilon_M .$$

In this section, we will examine models for the error in the ACS, the PEP, and the inconsistency in characteristics between the two.

An underlying assumption is that the ACS distribution within foreign-born of those reporting a residence outside the U.S. one year ago during January through December of the collection year is equal to the distribution for those who entered the U.S. July 1 of the year before the collection year to June 30 of the collection year. At this point in the development, we will assume the error due to any violation of this assumption is negligible.

3.1 Models for data error

The estimation of foreign-born immigration involves two sources of data, PEP and ACS. We examine models for the errors separately for the two data sources.

3.1.1 Model for ACS data error

First, focusing on the estimates from the ACS in terms of the rates $\hat{\delta}_F$ and $\hat{r}_Y$ appears helpful in deriving estimates of bias and random error. Collecting new data regarding the bias and random error is not possible with the current resources. So, the only available data are from previous ACS evaluations and ACS production information. Possibly, evaluation data from other surveys will also be helpful in deriving estimates. In this

situation, the information about the bias and random error probably will be in the form of error rates. Also, there probably will not be separate estimates of error rates for each ratio adjustment cell. Therefore, cells may have to be collapsed for the analysis.
For examining the bias and random error of the estimates $\bar{s}_F$ and $\bar{r}_Y$ from the ACS, consider the following form

$$\bar{s}_F = s_F + \beta_F + \varepsilon_F$$
$$\bar{r}_Y = r_Y + \beta_Y + \varepsilon_Y$$

where ε_F and ε_Y are random error terms such that $E[\varepsilon_F] = E[\varepsilon_Y] = 0$, and β_F and β_Y are biases.

Notice that the product of $\bar{s}_F$ and $\bar{r}_Y$ may be written as

$$\begin{aligned}\bar{r}_Y\bar{s}_F &= [(r]_Y + \beta_Y + \varepsilon_Y)[(s]_F + \beta_F + \varepsilon_F) \\ &= r_Y s_F + r_Y(\beta_F + \varepsilon_F) + s_F(\beta_Y + \varepsilon_Y) + (\beta_F + \varepsilon_F)(\beta_Y + \varepsilon_Y)\end{aligned}$$

It seems reasonable to assume that the last term is negligible since it is the product of errors in rates. Under that assumption, the $\bar{s}_F\bar{r}_Y$ would have negligible error if

$$r_Y(\beta_F + \varepsilon_F) = -s_F(\beta_Y + \varepsilon_Y)$$

If $(\beta_F + \varepsilon_F)$ and $(\beta_Y + \varepsilon_Y)$ are both nonzero, they would have to have opposite signs since both $\bar{s}_F$ and $\bar{r}_Y$ are positive.

Decompose the bias terms into bias from frame coverage error (*cov*), data collection error (*dc*), and data processing error (*dp*)

$$\beta_F = \beta_{F-cov} + \beta_{F-dc} + \beta_{F-dp}$$
$$\beta_Y = \beta_{Y-cov} + \beta_{Y-dc} + \beta_{Y-dp}$$

The random error may be expressed as variance with a term due to sampling error (*samp*) and a term due to imputation error (*imp*) where

$$\varepsilon_F^2 = V_F = V_{F-samp} + V_{F-imp}$$

$$\varepsilon_Y^2 = V_Y = V_{Y-samp} + V_{Y-imp}$$

The next step is to derive reasonable estimates of the bias and random error terms for $\bar{s}_F$ and $\bar{r}_Y$. Probably the best approach is to conduct a simulation to synthesize the estimates of error from the different sources. The reason is that estimates of the bias terms will have their own random error and a simulation approach could account for this.

For estimates of the bias terms, examining available data for the different aspects of the data collection and data processing operations may provide estimates for the bias components.

- For coverage error estimates, we will examine the available information about the coverage of housing units by the ACS sampling frame
- For the data collection error estimates, we will examine the available information on roster errors, errors in the designation of foreign-born, errors in the designation of residing outside the U.S. one year ago for the foreign born, and other field errors, such as interviewers going to the wrong address.
- For data processing error estimates, we will examine the available information about errors in the editing, coding, and keying relevant to the designation of foreign-born and the designation of residence outside the U.S. one year ago for the foreign-born.

Random error usually is influenced by sampling error and imputation error. These may be considered separate variance components.

- Estimates of the sampling error may be obtained from the ACS. The sampling error may be obtained for the rates $\bar{y}_F$ and $\bar{y}_Y$.
- Another source of random error may arise from imputation for missing data. This may be in the form of error due to the choice of the model used by the ACS imputation or error due to estimating parameters in the chosen imputation model.

3.1.2 Model for PEP data error

The ratio adjustment of the ACS estimates to the estimates from PEP relies on the assumption that the PEP estimates of the total population in cells used in ratio adjustment are correct.

Sources of potential errors in estimates of population size from the PEP include:

- Coverage error in the Census 2000 numbers, which are used as a base
- Errors in the data used to form updated population estimates during the decade.

For examining the bias and random error the estimate $\bar{P}$ from the PEP for a ratio adjustment cell *C*, consider the following form

$$\bar{P} = P + \beta_P + \varepsilon_P$$

where ε_P is a random error terms such that $E[\varepsilon_Y] = 0$, and β_P is the bias.

Note that a contributor to the bias and random error terms for $\bar{P}$ for a particular year will be the foreign-born immigration estimates from ACS used in constructing immigration estimates in previous years. This is true for the PEP estimates of Vintage 2008 and later years. The migration estimates for PEP estimates of Vintages 2006 and earlier did not use the ROYA method, although Vintage 2007 used migration estimates that were partially based on the ROYA method.

3.2 Model for error from inconsistent characteristics between ACS and PEP

The ratio adjustment methodology relies on the assumption that the PEP and the ACS measure the characteristics used to form weighting cells in a consistent way.

Sources of inconsistencies in measurement of characteristics used in weighting cells between the PEP and ACS:

- Coding of responses of race and Hispanic ethnicity in the ACS
- Changes in reporting of race/Hispanic ethnicity since Census 2000, which is the base for the PEP
- Differences in characteristics used in forming ratio adjustment cells between the ACS and the data used to form updated PEP estimates during the decade. These differences could be caused by errors or changes in record keeping by the data sources.

There is some automated and clerical coding of write-in responses for race and Hispanic ethnicity. For quality assurance (QA), there is a sampling of batches. The acceptance sampling methodology assures that each sampled batch has an Average Outgoing Quality Limit (AOQL), or approximate resulting keying error of less than one percent for the whole form, not applied to individual questions (Wolfgang 2007). This tolerance is being raised to three percent beginning in 2010 (Wolfgang 2009).

To examine the model for the error from inconsistent characteristics, first let ratio adjustment cell C_k be the cell reported in the ACS while C_j denotes the person's cell according to how characteristics are reported for the PEP.

Then define the following:

$f(j,k)$ = the proportion of persons whose ACS characteristics put them in cell C_k while their PEP characteristics put them in cell C_j.

$f_F(j,k)$ = the proportion of foreign-born persons whose ACS characteristics put them in cell C_k while their PEP characteristics put them in cell C_j.

$f_Y(j,k)$ = the proportion of foreign-born persons whose ROYA was out of the country and whose ACS characteristics put them in cell C_k while their PEP characteristics put them in cell C_j.

Then we can construct estimates of Q, F, and Y for each ratio adjustment cell j that have the inconsistent ACS characteristics corrected to be consistent with the PEP characteristics as follows:

$$\bar{Q}_{j,I} = \sum_k f(j,k)\bar{Q}_k$$

$$\bar{F}_{j,I} = \sum_k f_F(j,k)\bar{F}_k$$

$$\hat{\bar{Y}}_{k,I} = \sum_{k} f_Y(j,k)\overline{Y_k} \quad .$$

Note that all the above summations include letting *k* equal *j*. Then the estimate of the foreign-born immigration for cell *j* using characteristics consistent with PEP is $\hat{T}_{j,I}$ defined by

$$\hat{T}_{j,I} = P\frac{\hat{F}_{j,I}}{\hat{Q}_{j,I}}\frac{\hat{Y}_{j,I}}{\hat{F}_{j,I}}$$

The bias in the ACS estimate for cell *k* is

$$\hat{B}_{k,I} = \hat{T}_k - \hat{T}_{k,I} \, .$$

Then we may form an estimate of the bias in the estimate of the total foreign-born immigration due to inconsistent characteristics

$$\hat{B}_I = \sum_{k} \hat{B}_{k,I} \quad .$$

We are not aware of any information about inconsistencies in reporting of characteristics between the PEP and the ACS. There is a study that compared the differences in characteristics for individuals on two occasions (Farber 2001). One of two occasions was an enumeration in Census 2000 and the other was an independent survey designed to measure census coverage error, the Accuracy and Coverage Evaluation Survey (A.C.E.). The study examined the inconsistencies in characteristics for people whose records in the census and the A.C.E. matched.

4. Indicators of general quality of ACS

Two indicators that are often used to judge the general quality of a survey are coverage ratios and response rates. A *coverage ratio* is calculated by dividing the estimated number of persons in a specific demographic group from the survey by an independent population total for that group (U.S. Census Bureau 2002, p. 16-1). For Census Bureau surveys, the independent estimates for calculating the coverage ratios are the PEP estimates. If a ratio adjustment to the PEP were done for the survey as a whole rather than by cells, the coverage ratio would be the reciprocal of the ratio adjustment factor.

The survey's *response rate* is the proportion of sample units that were eligible or of unknown eligibility that responded to the survey (expressed as a percentage). However, since ACS has three modes of interviewing and selects a subsample of the nonrespondents at the end of the first two modes, which are mail and Computer Assisted Telephone Interviewing (CATI), for the last mode of Computer Assisted Personal Interviewing (CAPI), the ACS response rate is calculated by

$$\text{ACS response rate} = \frac{\text{Weighted number of completed mail, CATI, \& CAPI interviews}}{\text{Weighted number of eligible cases}}$$

A possible indicator of quality is the patterns in the coverage ratios for ACS and its relative pattern when compared with other current surveys conducted by the Census Bureau. For example, if the coverage ratios for all the surveys are under 1, that may indicate undercoverage in the surveys, although it also could indicate problems with the independent estimate. If sampling error were the only thing the ratio adjustment was addressing, the coverage ratios would vary over the years, sometimes greater than 1 and sometimes less than 1.

Figures 1 to 4 contain coverage ratios from 2000 to 2008 for Blacks, Nonblacks, males, and females from ACS, Current Population Survey (CPS), National Crime Victimization Survey (NCVS), Survey of Income Program Participation (SIPP), and National Health Interview Survey (NHIS). The CED and CEQ are not included on this graph because some of their coverage ratios are below 0.75, much lower than the ACS coverage ratios.

Figure 5 contains the response rates from 2000 to 2008 for ACS and the other six surveys mentioned in the previous paragraph.

The coverage ratios and the response rates for ACS are higher than for the other surveys and appear less variable. However, to be fair, ACS has a longer period for data collection than the other surveys and a sequence of three modes of data collection, which are mail, telephone, and personal visit, where the other surveys do not.

The ACS coverage ratios for 2001 to 2008 appear fairly stable although a test for trend could be done for ACS and the other surveys. The coverage ratio for 2000 is higher than the others, but there may be some explanation for that since the 2000 version was a large-scale survey designed to demonstrate that ACS could replace the census long form. The 2001 through 2004 implementations of ACS were smaller scale. Then in 2005, the full implementation of ACS began.

The coverage ratios for nonblacks are higher than those for blacks. And, the coverage ratios for females are higher than those for males.

Approximately one-third of the estimated immigrants are Hispanic so the coverage rate for Hispanics is important. Table 2 shows ACS coverage ratios by race and Hispanic ethnicity for 2007 and 2008. In both years, the coverage ratio for Hispanics is lower than for nonHispanic whites, Asians, and American Indians & Alaska Natives, but higher than for nonHispanic blacks. In 2007 the coverage ratio for Hispanics is lower than for Native Hawaiians & Pacific Islanders, but higher in 2008. However, the coverage ratio for Hispanics is fairly stable for these two years. This is true for the other race groups with the exception of the Native Hawaiians and Pacific Islanders, but there may be some explanation for the variation for this group.

Overall, the ACS coverage ratios appear stable, which is helped by the ACS having a high response rate. This suggests that the PEP estimate and the unadjusted ACS estimate tend to have similar percentage year-to-year changes.

Table 2. Coverage ratios for 2007 and 2008 ACS by race/Hispanic ethnicity

	2007	2008
Total:	94.2	93.8
Not Hispanic or Latino:		
White	95.4	94.7
Black or African American	89.1	89.7
American Indian and Alaska Native	96.8	96.2
Asian	95.6	96.9
Native Hawaiian and Pacific Islander	96.1	85.8
Hispanic or Latino	92.8	92.5

Source: American Fact Finder. Table B98013.

5. Sources of potential errors in ACS

In this section, we explore models and the information available to form estimates of the errors in the estimated proportions $\bar{s}_F$ and $\bar{r}_Y$ in a ratio adjustment cell *C* (see Section 3.1) arising during the data collection. The error terms we examine arise from coverage errors β_{F-cov} and β_{Y-cov}, data collection errors β_{F-dc} and β_{Y-dc}, and data processing errors β_{F-dp} and β_{Y-dp}.

5.1 Frame Errors

Frame errors are errors in construction of the list used for sampling that may result in coverage error of the population. There are two types of error related to living quarters:

- Housing Unit (HU) frame errors: (1) omissions; (2) erroneous inclusions: no living quarters, duplicate of other HU or GQ; (3) misclassification of HU as GQ.
- Group Quarters (GQ) frame errors: (1) omissions; (2) erroneous inclusions: no living quarters, duplicate of other HU or GQ; (3) misclassification of HU as GQ

One source of information about the quality of the Master Address File (MAF) comes from the Frame Assessment for Current Household Surveys (Li, Loudermilk, and Liu 2008). This assessment was performed for household surveys the Census Bureau conducts other than ACS in preparation for the redesign of these surveys in 2014. All use the centralized MAF but have somewhat different filtering criteria for considering an address to be residential. In the study, interviewers went to the blocks with lists of addresses from the MAF. They determined whether an address was residential and added residential addresses that were not on the MAF. The data was collected before the block canvassing operation for the 2010 Census.

Nationwide, the study found that the omission rate for HUs was 4.67 percent. An additional 1.73 percent were on the MAF but were filtered out in creating the address list.

The study produced an interesting result for mobile homes, which are 3.28 percent of all housing units. Of the total number of mobile homes at the end of the study, 15.15 percent had not been on the MAF.

These results have to be viewed with some caution because at least some of the HUs found may not have been occupied. If a HU is vacant, missing it will not affect the estimates for people. This study was conducted prior to the Address Canvassing operation in 2009 for the 2010 Census, the results of which improved the MAF (Kennel and Martin 2010). However, since the focus of the paper is for estimates created with data collected prior to Address Canvassing so its improvements would not be relevant.

We will use this information to form the estimates of the bias terms $\tilde{\beta}_{F-cov}$ and $\tilde{\beta}_{Y-cov}$. We may collapse ratio adjustment cells to form larger groups for the estimation.

5.2 ACS Data Collection Errors

Errors may occur when the respondent fills out the mail questionnaire, or errors may occur during the interaction between the respondent and interviewer during the interviews. A particular mode of data collection may be more prone to specific errors than the other modes. This is important to explore because many of the interviews with foreign-born respondents tend be by telephone or in person via CATI or CAPI, respectively.

Potential sources of ACS data collection errors that may contribute to the data collection error terms β_{F-dc} and β_{Y-dc} in the estimated proportions $\hat{p}_F$ and $\hat{p}_Y$, respectively, in a ratio adjustment cell *C* (see Section 3.1) fall into four categories:

- Net error in whether foreign-born or native-born
- Error among foreign born that leads to error in whether lived outside U.S. one year ago
- Address errors
- Roster errors

<u>Error in whether foreign-born or native-born</u>
Cognitive testing and field testing have shown that respondents appear to understand the question about citizenship very well (Harris et al 2007). Foreign-born individuals who are now naturalized U.S. citizens remember very well the day they and other family members became U.S. citizens. There was some confusion about the category "born outside the U.S. to U.S. parents," particularly when the respondent was not one of the parents. Although these children are considered to be "native-born", sometimes the answer for them was "foreign-born."

The cognitive interviews did not observe errors in other categories. A very small portion of the population falls in the category "born outside the U.S. to U.S. parents" where there were some errors.

Another source of information about error in responses to a question about citizenship comes from a response variance study for the National Survey of College Graduates (NSCG) (Singer 2005). The NSCG asks whether the respondent was a U.S. citizen or not a U.S. citizen. All the responses are self responses. There was almost no variance in the responses for this version of a citizenship question and this limited population.

These studies suggest that any error introduced in the estimated proportion $\hat{\delta}_F$ has to be very small. We will define β_{F-q} to be the net error in the proportion $\hat{\delta}_F$ caused by error in response to the question.

Error among foreign born that leads to error in whether lived outside U.S. one year ago
Not much information exists about the quality of reporting by the foreign-born on whether they lived outside the U.S. one year ago. There was some field testing of two versions of the question in the 2006 American Community Survey Content Test. This test focused on how well respondents provided specific details of their previous address for the geographic coding. There was not enough foreign-born in sample to make statistical inferences.

We will define β_{Y-q} to be the net error in the proportions $\hat{r}_Y$ caused by error in response to the question.

Address errors
Address errors cause interviews to be mistakenly conducted at an address other than the sample address. This includes mail delivery mix-ups as well as interviewers going to the wrong address. This category also includes misclassification errors regarding whether an address is a housing unit or not and whether a housing unit is vacant or occupied.

A source of information about address errors comes from the ACS Reinterview (Peterson 2008). The major finding regarding address errors was that some refusals were classified as vacant. We are examining the data to determine whether the tracts where these misclassifications occurred are different from the whole population in any way relevant to the immigration estimates.

We will define β_{F-a} to be the net bias in the proportion $\hat{\delta}_F$ due to address errors. We will define β_{Y-a} to be the net bias in the proportion $\hat{r}_Y$ due to address errors.

Roster errors
There are two basic types of roster errors: (1) incorrectly excluding a household member; and (2) incorrectly including a household member.

A source of information about roster errors comes from the ACS Reinterview (Peterson 2008). The study found that 98.7 percent of the Reinterview rosters agreed completely with the ACS rosters. However, 129 ACS rosters had omissions and 107 rosters had erroneous inclusions. We are examining the data to determine whether the tracts where these roster errors occurred are different from the whole population in any way relevant to the immigration estimates.

We will define β_{F-r} to be the net bias in the proportion $\bar{s}_F$ due to roster errors. We will define β_{Y-r} to be the net bias in the proportion $\bar{r}_Y$ due to roster errors.

Combining estimates of potential sources

The data collection error β_{F-dc} is the sum of the three terms

$$\beta_{F-dc} = \beta_{F-q} + \beta_{F-a} + \beta_{F-r}.$$

The data collection error β_{Y-dc} is the sum of the three terms

$$\beta_{Y-dc} = \beta_{Y-q} + \beta_{Y-a} + \beta_{Y-r}.$$

To form the estimates of the terms in the bias components $\hat{\beta}_{F-dc}$ and $\hat{\beta}_{Y-dc}$, we need to consider that there are multi-mode aspect of ACS because many of the interviews of foreign-born respondents tend to occur by telephone or in person. For a cell C, we take an approach similar to the one in Section 2 and define the following:

$\bar{Q}_m$ = sum of the weights of those who responded by mail

$\bar{Q}_t$ = sum of the weights of those who responded by telephone

$\bar{Q}_p$ = sum of the weights of those who responded in person.

$\bar{F}_m$ = sum of the weights of those who responded by mail and are foreign-born

$\bar{F}_t$ = sum of the weights of those who responded by telephone and are foreign-born

$\bar{F}_p$ = sum of the weights of those who responded in person and are foreign-born.

Then, we can write an estimate as the weighted sum of the estimates for each response mode where the weights are the weighted proportions for the response mode. An example follows for the proportion of the population who are foreign-born.

$$\begin{aligned} \bar{s}_F &= \frac{\bar{F}_m + \bar{F}_t + \bar{F}_p}{\bar{Q}} \\ &= \frac{\bar{Q}_m}{\bar{Q}}\frac{\bar{F}_m}{\bar{Q}_m} + \frac{\bar{Q}_t}{\bar{Q}}\frac{\bar{F}_t}{\bar{Q}_t} + \frac{\bar{Q}_p}{\bar{Q}}\frac{\bar{F}_p}{\bar{Q}_p} \\ &= \frac{\bar{Q}_m}{\bar{Q}}s_{Fm} + \frac{\bar{Q}_t}{\bar{Q}}s_{Ft} + \frac{\bar{Q}_p}{\bar{Q}}s_{Fp}. \end{aligned} \qquad (1)$$

In case we find that we need to take the response mode into consideration when estimating the bias due to data collection error β_{F-dc}, we derive an estimator using Equation (1). For the derivation, we assume that the whole population answers the ACS questionnaire and that errors only occur during the data collection. Under these assumptions there is no random error $\bar{Q} = Q, \bar{Q}_m = Q_m, \bar{Q}_t = Q_t$, and $\bar{Q}_p = Q_p$. In addition,

since there is only one source of bias, $\hat{s}_F - s_F + \beta_{F-dc}$. Next let β_{Fm-dc} be the bias in proportion of foreign-born among those who respond by mail $\hat{s}_{Fm}$, β_{Ft-dc} be the bias in proportion of foreign-born among those who respond by mail $\hat{s}_{Ft}$, and β_{Fp-dc} be the bias in proportion of foreign-born among those who respond in person $\hat{s}_{Fp}$. Then using Equation (1),

$$s_F = \frac{Q_m}{Q}[(s]_{Fm} + \beta_{Fm-dc}) + \frac{Q_t}{Q}[(s]_{Ft} + \beta_{Ft-dc}) + \frac{Q_p}{Q}[(s]_{Fp} + \beta_{Fp-dc}) \qquad = s_F + \frac{Q_m}{Q}\beta_{Fm-dc} + \frac{Q_t}{Q}\beta_{Ft}$$

Therefore,

$$\beta_{F-dc} = \frac{Q_m}{Q}\beta_{Fm-dc} + \frac{Q_t}{Q}\beta_{Ft-dc} + \frac{Q_p}{Q}\beta_{Fp-dc}.$$

We can use this result to form the following estimator since ACS is conducted on a sample basis

$$\hat{\beta}_{F-dc} = \frac{\hat{Q}_m}{\hat{Q}}\hat{\beta}_{Fm-dc} + \frac{\hat{Q}_t}{\hat{Q}}\hat{\beta}_{Ft-dc} + \frac{\hat{Q}_p}{\hat{Q}}\hat{\beta}_{Fp-dc}. \qquad (2)$$

A similar derivation that results in the following estimator for the bias due to data collection error in the estimate of the proportion $\hat{r}_Y$.

$$\hat{\beta}_{Y-dc} = \frac{\hat{r}_m}{\hat{r}}\hat{\beta}_{Ym-dc} + \frac{\hat{r}_t}{\hat{r}}\hat{\beta}_{Yt-dc} + \frac{\hat{r}_p}{\hat{r}}\hat{\beta}_{Yp-dc}. \qquad (3)$$

In Equation (3), $\hat{\beta}_{Ym-dc}$ is the estimator of the bias in the proportion who resided outside the U.S. one year ago among the foreign-born who respond by mail $\hat{r}_{Ym}$, $\hat{\beta}_{Yt\ \ dc}$ is the estimator of the bias in the proportion who resided outside the U.S. one year ago among the foreign-born who respond by telephone $\hat{r}_{Yt}$, and $\hat{\beta}_{Yp-dc}$ is the estimator of the bias in proportion who resided outside the U.S. one year ago among the foreign-born who respond in person $\hat{r}_{Yp}$.

5.3 ACS Data Processing Errors

Data processing errors refer to errors in the editing, coding, and keying operations that lead to the wrong classification of a person on a survey roster. A particular mode of data collection may require more editing than the others and therefore create more opportunities for errors. Also, a particular mode of data collection may result in more coding errors than the others.

Potential sources of ACS data processing errors that may contribute to the data processing error terms β_{F-dp} and β_{Y-dp} in the estimated proportions $\hat{s}_F$ and $\hat{r}_Y$ in a ratio adjustment cell *C* (see Section 3.1) fall into three categories:

- Errors in editing
- Errors in coding
- Errors in keying

These types of errors may occur for responses. When questions left blank are imputed, any error is considered imputation error.

Editing errors

Editing errors that have the potential for affecting the immigration estimates are: (1) in whether foreign-born or native-born; (2) in whether lived outside U.S. one year ago among foreign-born.

We will define β_{F-e} to be the net bias in the proportion $\hat{P}_F$ due to editing errors. We will define β_{Y-e} to be the net bias in the proportion $\hat{P}_Y$ due to editing errors.

Coding errors

There is no coding required for designating a person on the ACS roster as being foreign-born because the answers to Question 8 are only check boxes. However, for a person to be considered as residing outside the U.S. one year ago, the check box for Question 15a has to be marked to indicate living outside the U.S. and the country where the person lived has to be written in. The write-in responses for one's residence one year ago are coded so they can be converted to electronic format. The geographic coding system first attempts an automated coding of the write-in response. The responses not coded by the automated system are referred to clerical coding.

We will define β_{Y-g} to be the net bias in the proportion $\hat{P}_Y$ due to geographic coding errors that cause a foreign-born person to be designated as residing outside the U.S. one year ago when really were in the U.S., and vice versa.

Keying errors

Keying error may occur when write-in responses on paper forms are converted to electronic format through keying. This error also only has the potential for affecting whether a person on the ACS roster is designated as living outside the U.S. one year ago.

ACS paper forms are keyed from an image obtained by scanning. The forms are double keyed with adjudication. The adjudication procedure makes the keying error very low as long as the scans of the forms are easy to read.

We will define β_{Y-k} to be the net bias in the proportion $\hat{P}_Y$ due to keying errors that cause a foreign-born person to be designated as residing outside the U.S. one year ago when really were in the U.S., and vice versa.

Combining estimates of potential sources

The data processing error β_{F-dp} has only one source, so

$$\beta_{F-dp} = \beta_{F-e}.$$

The data processing error β_{Y-dc} potentially has three sources. It may be expressed as the sum of the three terms

$$\beta_{Y-dp} = \beta_{Y-e} + \beta_{Y-g} + \beta_{Y-k}.$$

To form the estimate of the bias term β_{Y-dp}, we will either form separate estimates of the individual terms or form a combined estimate. We also may collapse ratio adjustment cells to form larger groups for the estimation of both β_{F-dp} and β_{Y-dp}.

5.4 Random Error

Our strategy is to estimate the components for random error for the estimate of foreign-born immigration rather than separately for the random error terms for ε_F and ε_Y. Therefore, we will estimate terms for sampling error and for imputation error.

5.4.1 ACS Random Error

Sampling error is random error that occurs because of selecting a sample rather than interviewing the entire population. Table 4 shows the estimates of the standard error due to sampling for the foreign-born population whose residence one year ago was outside the U.S. before the ratio adjustment to the PEP estimates. The estimates of the foreign-born immigration in Table 4 are after the ratio adjustment.

Table 4. ACS estimates of foreign-born population whose residence one year ago was outside the U.S. and their standard errors (thousands).

Period (July 1 to June 30)	ACS year	estimate	standard error (ε_S)
1999-2000	2000	1,420	42.4
2000-2001	2001	1,421	35.2
2001-2002	2002	1,228	33.6
2002-2003	2003	1,025	28.8
2003-2004	2004	1,124	31.3
2004-2005	2005	1,188	17.8
2005-2006	2006	1,190	15.9
2006-2007	2007	1,114	20.3
2007-2008	2008	1,082	16.8

Source: ACS
Note: Standard errors calculated with weights prior to population controls.

5.4.2 ACS Imputation Error

Every survey has missing data. Errors may occur during the course of the application of imputation methods to compensate for missing data. We will model imputation error as a random error component.
The imputation method that ACS uses to account for sample units that are not interviewed is a weight adjustment. ACS uses a geography-based hot-deck method to impute a value when there is missing data for whether foreign-born or native-born and when there is missing data for a person's residence one year ago.

There has not been a study of the random error introduced into ACS estimates through imputation. However, Kim, Fuller, and Bell (2008) have studied the sampling variance component due to the nearest neighbor imputation method for the 2000 Census long form. The ACS was designed to replace the long form so the questions are almost the same. Although the ACS imputation method uses the hot-deck method, it has some similarity to nearest neighbor imputation because the cells are drawn geographically so that ideally the donor is from the same block as the recipient. The study found that the imputation method increased the standard errors in the estimates of poverty and median income for the state of Delaware and Michigan by 16 to 30 percent.

A method for estimating the variance component due to choice of the imputation model involves constructing reasonable alternative imputation models and applying each model to construct an alternate set of imputations. This methodology has been used in error analyses for estimates of net coverage error in Census 2000 (Mulry and Spencer 2001, Mulry, ZuWallack, and Spencer 2003).

If there are L alternate models, an estimate of the foreign-born population, say $\hat{T}_j$, may be constructed using the data for the j^{th} model, where $1 \le j \le L$.

Then we estimate the variance due to imputation model $\widehat{V_M}$ by

$$\widehat{V_M} = \frac{\sum_{j=1}^{L} \left([\hat{T}_j - \overline{T_M}] \right]^2}{(L-1)}$$

where $\overline{T_M} = \frac{\sum_{j=1}^{L} \hat{T}_j}{L}$.

If a set of reasonable alternative imputation models are not available, an alternate approach is to perform a sensitivity analysis to explore the potential impact of the choice of the imputation model on the estimates of foreign-born immigration.

6. Simulation

Once the nonsampling errors, their variances, and covariance matrix are estimated, the simulation will draw repeatedly and independently from their joint distribution to produce the distribution of a bias estimate. The probability distributions will be centered on the observed values adjusted for the estimated biases, and their random component will be derived from a multivariate normal specification with mean vector equal to zero and covariance matrix. Simulation from this distribution will yield distributions of the estimates of immigration. Differences between the mean of the latter distribution and the original estimate indicate the estimated biases in the original estimates, and the standard deviations indicate the standard deviations of the sampling distributions.

As shown in the discussion of the joint error distribution, the probability models will be developed somewhat differently for (1) sampling error, (2) error from missing data, (3) effect of inconsistent classification, and (4) other frame coverage errors, data collection errors, and data processing errors.

7. Analysis of results

Part of the research will be to specify the domains C for the analysis. The following statistics will be computed from the simulated distribution: (i) estimate of bias, (ii) estimate of standard deviation (reflecting both sampling error and random nonsampling errors), and (iii) deciles of the distribution.

The analyses will include a sensitivity analysis to aid in determining the most influential error sources and error types.

References

Boertlein, C. G. and Peterson, A. K. (2007) "Evaluation Report Covering Residence 1 Year Ago (Migration)" by January 3, 2007. 2006 American Community Survey Content Test Report P.3. U.S. Census Bureau. Washington, DC.

Farber, J. (2001) "Accuracy and Coverage Evaluation: Consistency of Post-Stratification Variables" by Farber. February 28, 2001. DSSD Census 2000 Procedures and Operations Memo B-10. U.S. Census Bureau. Washington, DC.

Harris, P., Bhaskar, R., Shook-Finucane, C. and Ericson, L. (2007) "Evaluation Report Covering Place of Birth, U.S. Citizenship, and Year of Arrival" by January 12, 2007. 2006 American Community Survey Content Test Report P.1. U.S. Census Bureau. Washington, DC.

Kennel, T. and Martin, J. (2010) "Summary Report for Evaluating MAF Content Quality after the 2010 Decennial Address Canvassing." Doc. #2010-4.0-G-16, Version 1.0. Demographic Statistical Methods Division. U.S. Census Bureau. Washington, DC.

Kim, J., Fuller, W., and Bell, W. "Variance Estimation for the Nearest Neighbor Imputation for the U.S. Census Long Form." SRD Research Report Series #2008 -13. December 30, 2008. U.S. Census Bureau. Washington, DC.

Landman, C. (2009) "Updated Administrative Data and Performance Measures for Selected Demographic Survey – 2000 through 2008 REVISED with Updated ACS Data." Memorandum dated November 23, 2009. Demographic Surveys Division. U.S. Census Bureau. Washington, DC.

Li, M., Loudermilk, C., and Liu, X. (2008) "Frame Assessment for Curent Household Surveys (FACHS) National Evaluation: Final Analytic Report. Comparing a National Sample of Block Listings to the Master Address File" Doc. #2010-4.0-G-12, Version 1.0. Demographic Statistical Methods Division. U.S. Census Bureau. Washington, DC.

Mulry, M. H., ZuWallack, R., and Spencer, B. D. (2003). "Loss Function Analysis for A.C.E. Revision II Estimates of Census Coverage Error." *Proceeding of the Joint Statistical Meetings*. American Statistical Association. Alexandria, VA. 2966-2971.

Mulry, M. H. and Spencer, B. D. (2001) "Overview of Total Error Modeling and Loss Function Analysis.' DSSD Census 2000 Procedures and Operations Memorandum Series B-19*. U.S. Census Bureau. Washington, DC. http://www.census.gov/dmd/www/ReportRec.htm

Peterson, S. (2008) "Reinterview Results from the 2006 American Community Survey – Housing Unit." dated April 14, 2008. DSSD American Community Survey Memorandum Series RI-2007-04. U.S. Census Bureau. Washington, DC.

Singer, P. (2005) "Response Variance in the 2003 National Survey of College Graduates." RE Report 12. Demographic Statistical Methods Division. U.S. Census Bureau. Washington, DC.

U.S. Census Bureau (2002) "Current Population Survey Design and Methodology." Technical Paper 63RV. U. S. Census Bureau. Washington, DC.

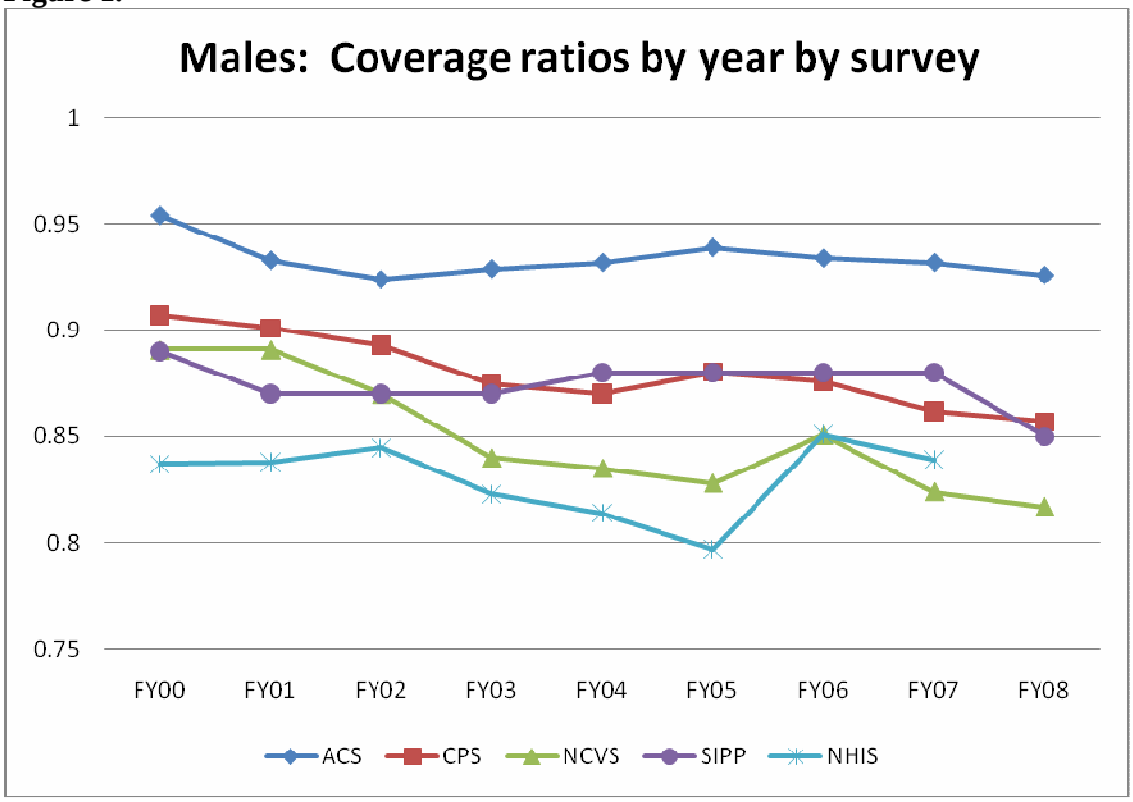
U.S. Census Bureau (2009) "Design and Methodology American Community Survey." Report ACS-DM1. U. S. Census Bureau. Washington, DC.

U.S. Census Bureau (2010) "Measuring net Internation Migration (NIM): Residence One Year Ago Method." Internal document dated January 14, 2010. Population Division. U.S. Census Bureau. Washington, DC.

Wolfgang, G. (2007) "Quality Assurance Specifications for the American Community Survey Key-From-Image Operation" DSSD AMERICAN COMMUNITY SURVEY MEMORANDUM SERIES QC 2007-01. U.S. Census Bureau. Washington, DC.

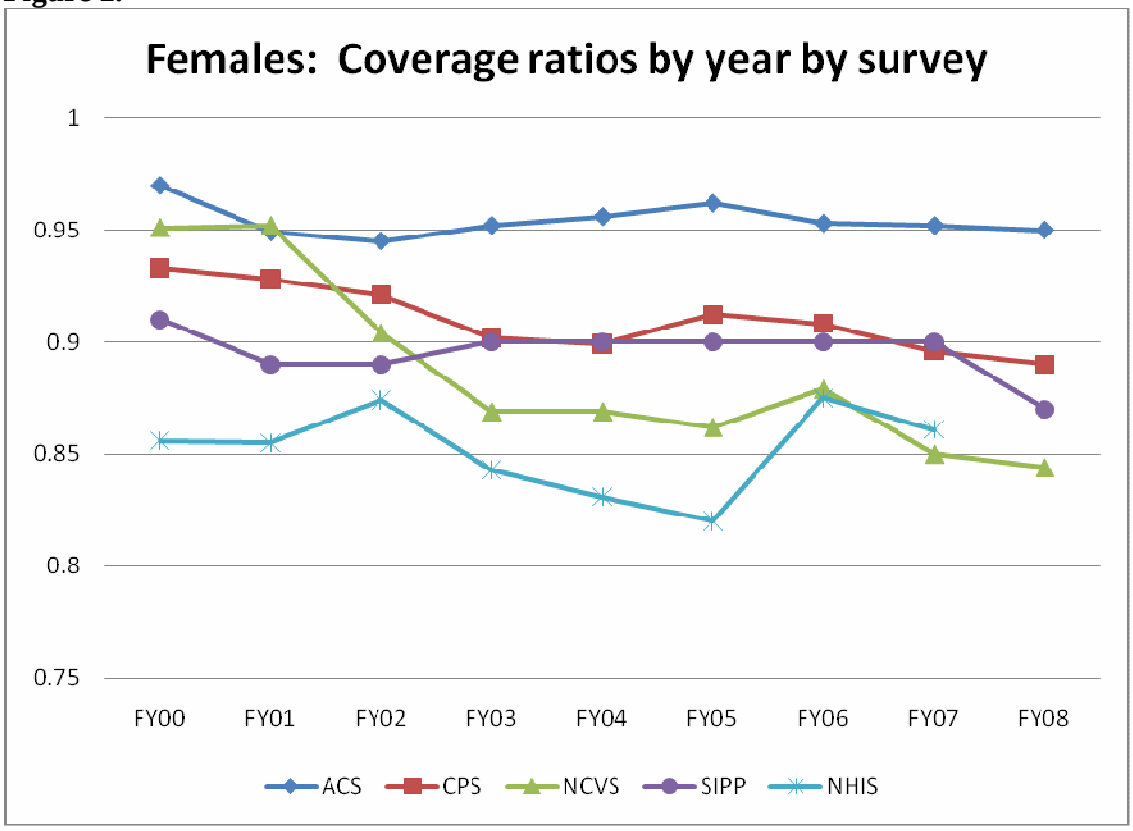
Wolfgang, G. (2010) "Quality Control Specifications for Response Coding of Race, Hispanic Origin, and Ancestry Write-In Responses for the American Community Survey." DSSD AMERICAN COMMUNITY SURVEY MEMORANDUM SERIES #QC 2009-02R. U.S. Census Bureau. Washington, DC.

Figure 1.



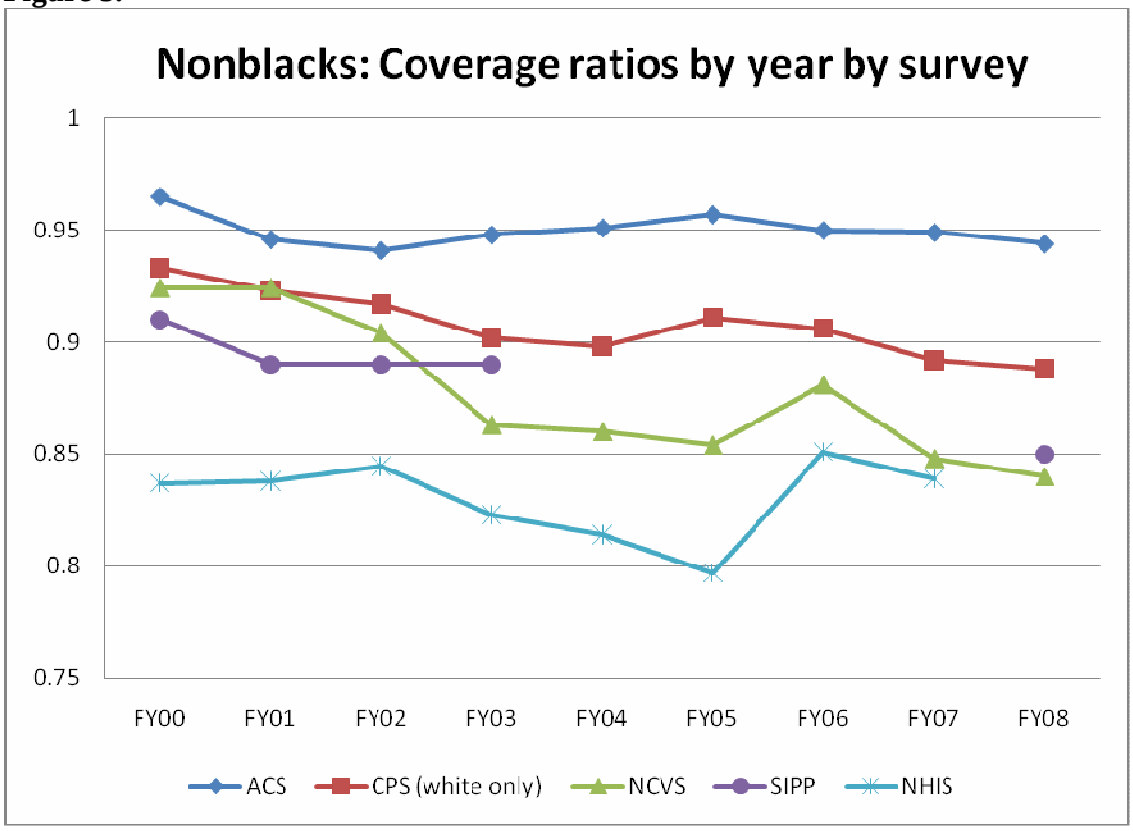
Source: (Landman 2009)

Figure 2.



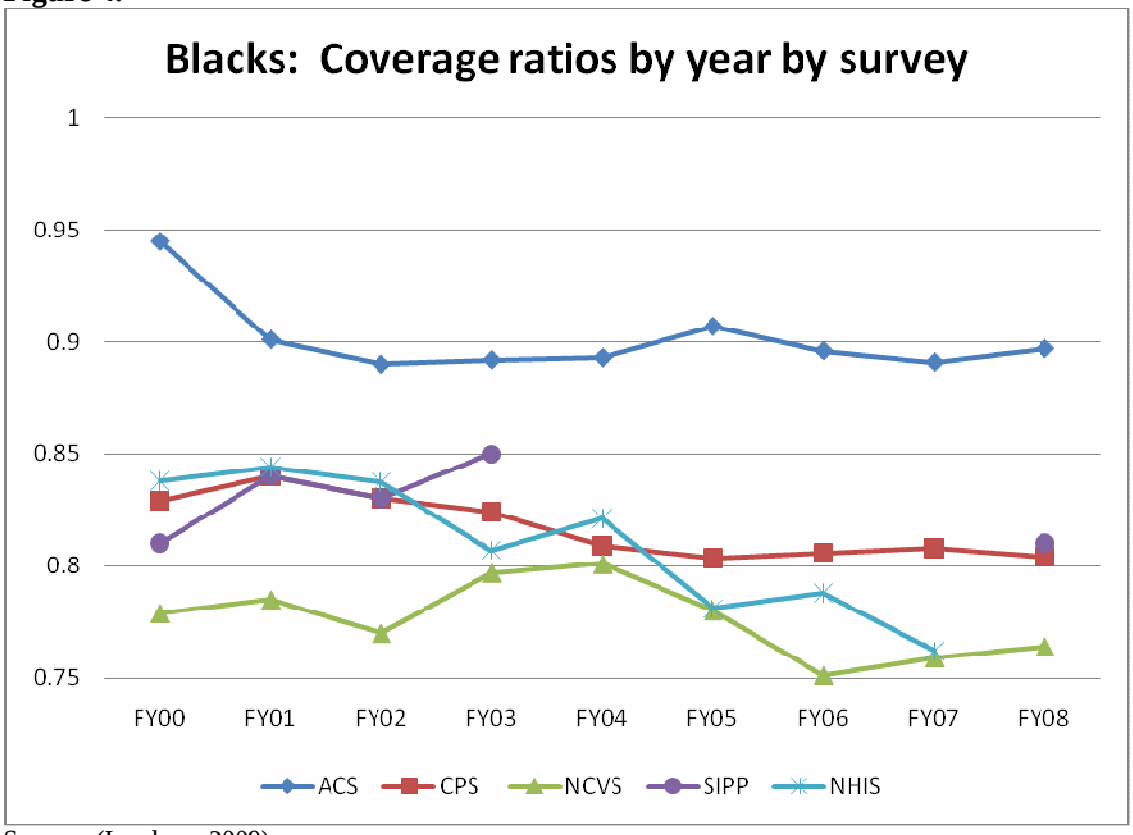
Source: (Landman 2009)

Figure 3.



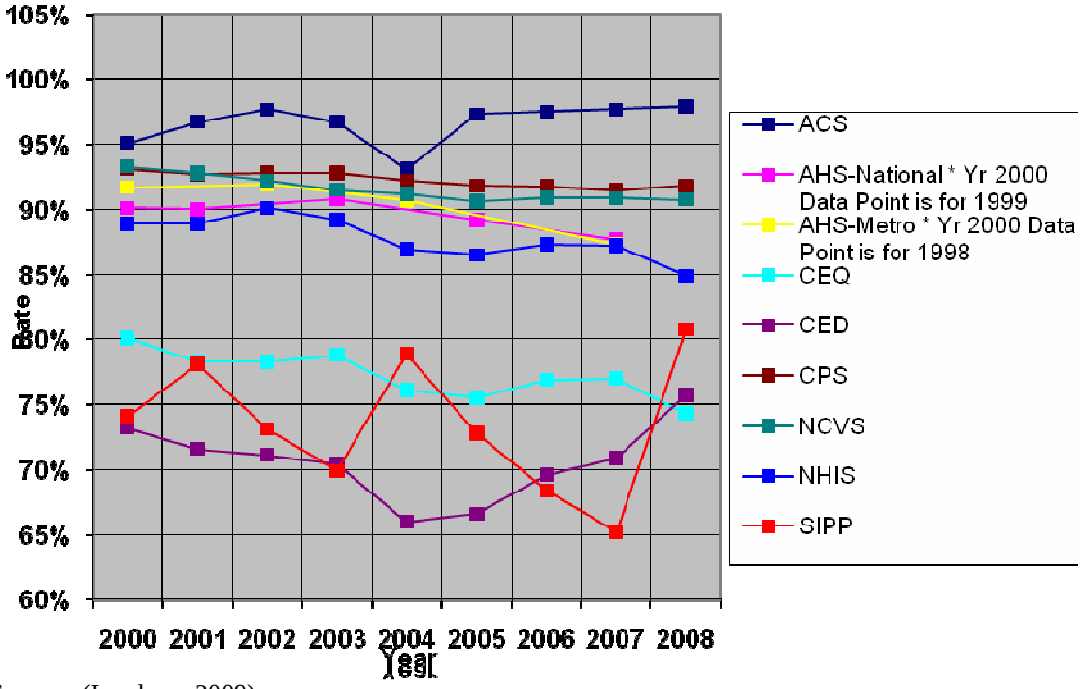
Source: (Landman 2009)

Figure 4.



Source: (Landman 2009)

Figure 5. Annual Response Rates: 2000-2008



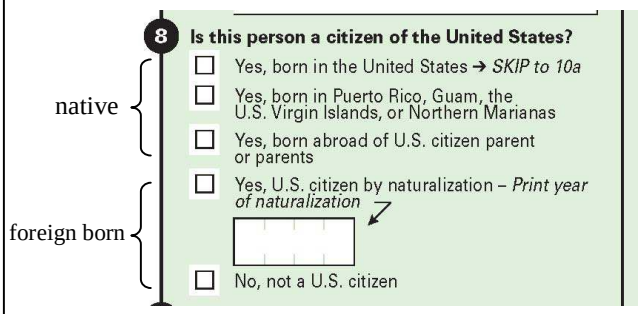
Source: (Landman 2009)

Appendix 1. ACS Questions for Foreign-born Immigration Estimation

ACS Question Number 8: Citizenship

"Is this person a citizen of the United States?"

- ❑ Yes, U.S. citizen by naturalization – Print year of naturalization _____
- ❑ No, not a U.S citizen



Question 15a; Residence one year ago

"Did this person live in this house or apartment 1 year ago?"

- ❑ No, outside the United States and Puerto Rico – Print name of foreign country, or U.S. Virgin Islands, Guam, etc. below _____