Statistical Journal of the IAOS 34 (2018) 179–189
DOI 10.3233/SJI-170374
IOS Press

When race and Hispanic origin reporting are discrepant across administrative records and third party sources: Exploring methods to assign responses[1]

Sharon R. Ennis*, Sonya R. Porter, James M. Noon and Ellen Zapata
Center for Administrative Records Research and Applications, U.S. Census Bureau, Washington, DC, USA

Abstract. The U.S. Census Bureau is researching uses of administrative records and third party data in survey and decennial census operations. One potential use of administrative records is to utilize these data when race and Hispanic origin responses are missing. When federal and third party administrative records are compiled, race and Hispanic origin responses are not always the same for an individual across sources. We explore different methods to assign one race and one Hispanic response when these responses are discrepant. We also describe the characteristics of individuals with matching, non-matching, and missing race and Hispanic origin data by demographic, household, and contextual variables. We find that minorities, especially Hispanics, are more likely to have non-matching Hispanic origin and race responses in administrative records and third party data compared to the 2010 Census. Minority groups and individuals ages 0-17 are more likely to have missing race or Hispanic origin data in administrative records and third party data. Larger households tend to have more missing race data in administrative records and third party data than smaller households.

Keywords: Race, Hispanic origin, administrative records, record linkage

1. Introduction

The U.S. Census Bureau is researching uses of administrative records and third party data (ARTPD) in survey and decennial operations in order to reduce costs and respondent burden while preserving data quality. One potential application of administrative records is to utilize the data when race and Hispanic origin responses are missing.

Item nonresponse for race and Hispanic origin is relatively low in census data. However, when a respondent does not provide a race or Hispanic origin, the Census Bureau employs methods such as hot decks to impute a response. A hot deck is geographically based, where responses from a nearest neighbor are used to impute missing responses to people with similar characteristics. The underlying assumption of a nearest neighbor hot deck is that people who live near each other share similar characteristics; however, with increasing racial and ethnic diversity in the U.S., this is less likely to be true [1].

For the first time in the 2010 Census, information that people had previously provided in either Census 2000 or the 2001–2009 American Community Surveys (ACS) were used to impute missing race and Hispanic origin responses. Previous census responses were used in almost 40 percent of all 2010 Census imputed His-

*Corresponding author: Sharon R. Ennis, Center for Administrative Records Research and Applications, U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233, USA. Tel.: +1 301 763 6041; E-mail: sharon.r.ennis@census.gov.

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

panic origin responses and 30 percent of all imputed race responses [2]. We may be able to expand on this imputation method to include other federal and third party sources [3].

Race and Hispanic origin responses in ARTPD may be able to assist not only with item nonresponse but also housing unit nonresponse. The Census Bureau is researching ways in which ARTPD could be used in decennial census operations when households do not respond to initial contact attempts. The quality of race and Hispanic origin data, as well as other demographic and housing data, in ARTPD is of crucial importance to this research.

However, when ARTPD are compiled, race and Hispanic origin responses are not always the same for an individual across different sources. In this paper, we explore different methods used to assign a single race and Hispanic origin response from ARTPD and evaluate which methods result in the highest level of agreement between an ARTPD composite of race and Hispanic origin responses and 2010 Census responses. We also describe the characteristics of individuals whose race or Hispanic origin responses in ARTPD match or do not match 2010 Census data, or have missing race or Hispanic origin responses in ARTPD.

In the next sections of this paper, we provide background on previous research on race and Hispanic origin data in ARTPD. Then we discuss the data and methods used in our analysis and present the results from our study. We conclude with a summary of our findings and propose future research.

2. Background

2.1. Census Bureau research on the quality of race and Hispanic origin in administrative records and third party data

In response to expanding interest in the use of administrative records to enhance a decennial census, the Census Bureau developed the Statistical Administrative Records System (StARS) in 1999. StARS 1999 was developed to support the Administrative Records Experiment which simulated Census 2000 counts with administrative records [4,5]. This previous research found that StARS had a lower representation of minorities compared to Census 2000. One of the limitations of the StARS administrative data was the inconsistent collection of race and ethnicity data. In particular, the Social Security Administration Numerical Identification File (Numident), which provided the widest coverage of race and ethnicity for the population, included Hispanic as a race category and did not collect multiple race responses [5]. In contrast, census data, adhering to the 1997 Office of Management and Budget (OMB) race and ethnicity standards, collects race and Hispanic origin as separate questions and starting in 2000 allowed for multiple-race reporting. StARS also modeled race and Hispanic origin when this information was missing in StARS, which likely contributed to differences between Census 2000 and StARS race and Hispanic origin data.

In a more recent study, the "2010 Census Match Study," Rastogi and O'Hara [6] expanded on this research and evaluated the agreement of demographic responses in ARTPD compared to the 2010 Census. In addition to the administrative sources used in StARS, this study utilized thirteen additional federal and third party files. Rastogi and O'Hara [6] found that non-Hispanics had higher agreement rates compared to Hispanics. Race response agreement varied by race group. The White alone, Black alone, and Asian alone populations had higher agreement rates compared to the Two or More Races, Native Hawaiian or Other Pacific Islander (NHPI) alone, American Indian or Alaska Native (AIAN) alone, and Some Other Race (SOR) alone populations. In a study that replicated the "2010 Census Match Study" using data from the 2010 ACS, Bhaskar et al. [7] found results for race and Hispanic origin that were consistent with those found by Rastogi and O'Hara [6].

The race and Hispanic origin agreement patterns observed by Rastogi and O'Hara [6] and Bhaskar et al. [7] are consistent with literature on racial and ethnic fluidity.

2.2. Racial and ethnic fluidity

One reason an individual's race or Hispanic origin in administrative records may not match their response in census data is racial and Hispanic origin fluidity. Individuals may change their identity and/or identification over time or in different situations and contexts (e.g., see [8–10]). Race response change varies considerably by race group. Non-Hispanic Whites, Blacks, and Asians are usually consistent in their race responses; while race response change is more common among non-Hispanic AIAN, NHPI, and multiracial individuals [8,10–12]. Previous Census Bureau research from the 1990, 2000, and 2010 Censuses shows that individuals are relatively consistent in their responses to

the Hispanic origin question with three percent or less changing their answer between the census and its corresponding reinterview [13–15].

Prior research shows substantial racial fluidity among Hispanics relative to non-Hispanics [10,13,14, 16]. One factor that may affect race reporting among Hispanics is that although the federal government defines race and ethnicity as separate concepts, many Hispanics view race and ethnicity as one concept and identify their race as "Hispanic." When faced with the federal standard racial categories, people who view their race as Hispanic may 1) not answer the race question, 2) report Hispanic responses that are tabulated as SOR, or 3) report a category that they feel may not be the best fit for their racial identity. Another factor affecting Hispanic racial identification is differences in questionnaire design. Campbell and Rogalin [17] conducted a study that compared responses from separate ethnicity and race questions to a combined ethnicity and race question for the same respondent. The authors found that most Hispanics who chose a race in the separate question identified as Hispanic only to the combined ethnicity and race question.

2.3. *Characteristics of people with non-matching and missing Hispanic origin and race responses*

Previous research on non-matching race and ethnicity data found that agreement varies by demographic and socioeconomic characteristics. Males [18, 19] and younger individuals [10,20] are more likely to have non-matching responses compared to females and older individuals. Individuals living in more affluent neighborhoods [18], people who live in the West, and people who respond through an interviewer compared to those who respond through mail tend to have non-matching race and Hispanic origin responses [10]. Household structure also has an impact on responses. Those living alone have more consistent responses than those living with others [21].

Few studies look at the patterns of missing race and Hispanic origin data in administrative records. We provide a brief overview of studies that have evaluated these patterns, but the findings are largely based on Medicaid [18,19] or Veteran's data [21] and may not apply to other ARTPD. Previous studies comparing survey data to administrative records found that White and younger individuals are more likely to have missing race responses [19,21]. However, Fernandez et al. [18] found that individuals who are Hispanic, AIAN, and older are more likely to have missing race responses in Medicaid administrative records. Males and people living in neighborhoods with higher median household incomes also tend to have missing race responses in Medicaid data [18]. Similar to patterns for missing race, minorities and people living in neighborhoods with higher median household incomes are more likely to have missing Hispanic origin responses in Medicaid data [18]. However, in contrast to missing race findings, females are more likely to have missing Hispanic origin responses in Medicaid data compared to males [18].

3. Data and methods

We used federal, state, and third party files to build a race and Hispanic origin ARTPD composite. We used previous census records (Census 2000 and ACS data from 2001 to 2009), Numident, Department of Housing and Urban Development (HUD) Tenant Rental Assistance Certification System (TRACS), HUD Public and Indian Housing Information Center (PIC), HUD Computerized Homes Underwriting Management System (CHUMS), the Center for Medicare and Medicaid Services Medicare Enrollment Database (MEDB), Indian Health Service (IHS) file, and Temporary Assistance for Needy Families (TANF) data in our race assignment methods as well as our regression analysis. We used these same files plus Texas Supplemental Nutrition Assistance Program data (SNAP), Medicaid Statistical Information System (MSIS), and third party files to assign Hispanic origin, as these additional sources indicated high levels of agreement for Hispanic origin responses but not race responses.

Administrative records sources vary in the collection of Hispanic origin and race data. Many of the federal files report race and ethnicity according to OMB's revised 1997 race and ethnic standards (see [22] for more information). However, there are a few exceptions. HUD TRACS collects an individual's ethnicity and race, but the 2010 HUD TRACS dataset used in this study has information for individuals for Hispanic origin or race, but not both. The Numident, MEDB, and Texas SNAP files treat race and Hispanic origin as one concept and have one combined race and ethnicity variable. In other words, the categories of the variable include "Hispanic" in addition to the race groups. Additionally, the Numident and MEDB data have a combined category for Asian and Pacific Islander and do not collect multiple responses or include a category for multiracial persons.

In order to compare the race and ethnicity data from the Numident, MEDB, and Texas SNAP files to the 2010 Census, we recoded the combined race and ethnicity variable into two separate variables, one for ethnicity and one for race. Individuals who were identified as Hispanic were coded as such with missing race information since we have no information about their race. Similarly, individuals who were identified as a race were coded as that race group with missing Hispanic origin information. For example, if an individual identified as Black, then the separate ethnicity variable was coded as missing and the race variable was coded as "Black." If an individual was identified as the combined category Asian/Pacific Islander then their race was coded as missing since we cannot determine with which OMB racial category – Asian or NHPI – the individual identifies.

Although the HUD PIC, HUD CHUMS, and TANF files collect race and Hispanic origin according to the OMB standards, these files do not include a category for SOR, unlike census data. The IHS file only identifies individuals as either AIAN or non-AIAN. The third party files model race and Hispanic origin data using information on surname and geography.

We link the ARTPD race and Hispanic origin composite to 2010 Census data for our analyses. All person records were processed through the Person Identification Validation System (PVS), which used probability record linkage techniques [23] and personal information such as name and date of birth to assign an anonymized unique Protected Identification Key (PIK) to each person, as possible (see [24]). The method is least robust for people who do not have a Social Security Number and those whose personal information is ambiguous or incomplete. Once the PIK was assigned in each separate data set, it was used to link a person's record in the 2010 Census to his or her own record in the race and Hispanic origin ARTPD composite.

We began with all individuals who were assigned a unique PIK in Census 2010 (268,706,490 records).[2] For the Hispanic origin analysis, we excluded individuals with edited Hispanic origin responses (11,967,175 records excluded). Similarly, we excluded individuals with edited race responses from the race analysis (9,211,212 records excluded). We excluded the non-Hispanic SOR and non-Hispanic multiracial groups from the analysis because not all ARTPD sources have an SOR category or collect multiple races (approximately 5.5 million records excluded from both the Hispanic origin and race analyses). Our data include 251,320,952 individuals in the Hispanic origin analysis and 253,905,696 in the race analysis.

We use descriptive statistics to evaluate two methods to assign Hispanic origin from ARTPD and six methods to assign race. We evaluate the methods by matching Hispanic and race response results from each method to 2010 Census unedited race and Hispanic origin responses.[3] Records without any available Hispanic origin or race data are not included in the descriptive match rates.

We chose one promising Hispanic origin assignment method and one promising race assignment method based on the descriptive statistics and applied multinomial regression analysis to understand the characteristics of those who have matching, non-matching, and missing race and Hispanic origin data. We perform multinomial regression analysis separately for Hispanic origin and for race. These models predict whether a linked Census-ARTPD record matches on Hispanic origin or race (coded as "0"), whether the Hispanic origin or race data do not match (coded as "1"), and whether the ARTPD record does not have any available Hispanic origin or race data (coded as "2"). Because the dependent variables include ARTPD records with missing demographic data, the distributions for the dependent variables differ from the distribution for matching Hispanic origin and race data presented in the descriptive analysis. As with the descriptive statistics, the models are limited to census records that are unedited.

The independent variables for the regressions include individual-level demographic variables, household-level characteristics, tract-level contextual characteristics, and region. Individual-level variables include the person's Hispanic origin, race, age, and gender as reported in the Census. We used a combined Hispanic origin and race independent variable with categories Hispanic, non-Hispanic White alone, non-Hispanic Black alone, etc. Household-level variables include household tenure, household type and size as reported in the Census, the Census mode in which the household responded, and whether the household lives in an urban or rural area. In addition, tract-level vari-

[2] Approximately 90 percent of 2010 Census records received a PIK (279,179,329 records). During the deduplication process, approximately 10 million records were removed from the data.

[3] We do not consider either source of data, the 2010 Census or the ARTPD, to be truth. We calculate the Hispanic origin and race agreement rates between the two sources without assigning greater value to either source.

Table 1
Percentage of records in the ARTPD composite with a discrepancy across source files or missing data

	Hispanic origin		Race	
	Number	Percent	Number	Percent
Total records	351,618,175	100.0	351,618,175	100.0
No discrepancy across source files	277,723,876	79.0	273,759,382	77.9
One source with data	86,447,226	24.6	90,743,297	25.8
Two or more sources with data	191,276,650	54.4	183,016,085	52.0
Discrepancy across source files	11,378,349	3.2	9,091,536	2.6
Missing data	62,515,950	17.8	68,767,257	19.6

Source: Administrative records and third party data. Note: ARTPD = Administrative records and third party data.

ables measure the percent of non-Hispanic Whites in the tract in the Census and the logged median household income in the tract according to ACS 2006–2010 5-year data.

3.1. Limitations

Our analysis does not include people in administrative records who were not assigned a PIK. The characteristics of individuals who receive a PIK are different from those who do not receive a PIK [25], which could bias our results. In addition, people in administrative records that did receive a PIK but could not be linked to 2010 Census data are not included in the analysis. This too is likely to result in some bias in our findings. Therefore, our results are not representative of the ARTPD population or the population enumerated in the 2010 Census and should be interpreted with caution.

4. Results

We first discuss Hispanic origin and race assignment methods and which methods resulted in the highest match between ARTPD and 2010 Census data for race and Hispanic origin responses. Then, we discuss results from the multinomial regression analysis.

4.1. Hispanic origin and race assignment methods

We developed methods to assign Hispanic origin and race data to the ARTPD composite based on available information in the administrative records and third party files. Once we assign a Hispanic origin or race to a record, the response is not overwritten by responses from any other files.

4.1.1. Hispanic origin

We considered two different methods in assigning a Hispanic origin response. Figure 1 illustrates how we applied these methods to assign one Hispanic origin response to the ARTPD composite. If there was no discrepancy in an individual's ethnicity response across source files then that response was assigned to the composite. This is shown in the first three rows of the figure. Of the 352 million records in the ARTPD composite, there were 278 million (79 percent) that had no discrepancy in Hispanic origin responses (see Table 1). Approximately 86 million (25 percent) had only one source of Hispanic origin data and 191 million (54 percent) records had the same Hispanic origin response across two or more sources.

There were 11 million (3 percent) individuals in the composite with discrepant ethnicity responses. When there are differences across the administrative records and third party source files, we assigned a Hispanic origin response according to the following:

In Method 1, a Hispanic response was assigned if a Hispanic response was present in any of the ARTPD sources. This is reflected in rows 4 and 5 in the "Method 1" column. ARTPD does not cover Hispanics as well as non-Hispanics and the agreement between ARTPD and census Hispanic responses is lower compared to non-Hispanics [6,7,26]. Thus, in this method, we gave priority to Hispanic responses to maximize the coverage of Hispanic responses in the ARTPD composite.

In Method 2, if a Hispanic origin response was found in previous census records, then that Hispanic or non-Hispanic response was assigned to the composite. This is shown in row 4 of the "Method 2" column in Fig. 1. Hispanic origin identification can be affected by questionnaire design [17]. Therefore, in this method, we gave priority to previous census records since the format is the most similar to the 2010 Census, relative to ARTPD sources. If a response was missing in previous census records, then a Hispanic response was assigned if present in any other administrative records source. This is reflected in row 5.

As shown in the last row of the figure, if Hispanic origin information is missing, then the response is set

Example #	ARTPD Source 1	ARTPD Source 2	ARTPD Source 3	Previous Census	**Method 1**	**Method 2**
1	Non-Hispanic	Non-Hispanic	Non-Hispanic	Non-Hispanic	**Non-Hispanic**	**Non-Hispanic**
2	Missing	Non-Hispanic	Missing	Missing	**Non-Hispanic**	**Non-Hispanic**
3	Hispanic	Hispanic	Missing	Hispanic	**Hispanic**	**Hispanic**
4	Hispanic	Hispanic	Missing	Non-Hispanic	**Hispanic**	**Non-Hispanic**
5	Non-Hispanic	Non-Hispanic	Hispanic	Missing	**Hispanic**	**Hispanic**
6	Missing	Missing	Missing	Missing	**Missing**	**Missing**

ARTPD = Administrative records and third party data.

Fig. 1. Methods used to assign Hispanic origin response to the ARTPD composite.

Example #	ARTPD Source 1	ARTPD Source 2	Previous Census	IHS	**Method 1**	**Method 2**	**Method 3**	**Method 4**	**Method 5**	**Method 6**
1	Missing	White alone	White alone	Missing	**White alone**	**White alone**	**White alone**	**White alone**	**White alone**	**White alone**
2	Black alone	Missing	Black alone	Missing	**Black alone**	**Black alone**	**Black alone**	**Black alone**	**Black alone**	**Black alone**
3	NHPI alone	Two or More	NHPI alone	Missing	**NHPI alone**	**Two or More**	**NHPI alone**	**NHPI alone**	**NHPI alone**	**NHPI alone**
4	NHPI alone	Missing	Missing	AIAN alone	**NHPI alone**	**NHPI alone**	**NHPI alone**	**NHPI alone**	**NHPI alone**	**AIAN alone**
5	AIAN alone	White alone	AIAN alone	AIAN alone	**AIAN alone**	**AIAN alone**	**AIAN alone**	**AIAN alone**	**AIAN alone**	**AIAN alone**
6	Asian alone	AIAN alone	AIAN alone	AIAN alone	**AIAN alone**	**AIAN alone**	**AIAN alone**	**AIAN alone**	**AIAN alone**	**AIAN alone**
7	Asian alone	Missing	Two or More	Missing	**Two or More**	**Two or More**	**Two or More**	**Two or More**	**Two or More**	**Two or More**
8	Missing	Missing	Missing	Missing	**Missing**	**Missing**	**Missing**	**Missing**	**Missing**	**Missing**

ARTPD = Administrative records and third party data; IHS = Indian Health Service.

Fig. 2. Methods used to assign race response to the ARTPD composite.

to missing in the ARTPD composite. Table 1 shows that approximately 63 million (18 percent) individuals in the ARTPD composite had no Hispanic origin information.

Once a single Hispanic origin response was assigned to the ARTPD composite using each method, the composite was then linked to 2010 Census data, and match rates for response agreement were calculated to evaluate the quality of the Hispanic origin responses in ARTPD. The match rates for each method are shown in Table 2. The agreement rate for Hispanics in Method 1 is higher (94 percent) compared to Method 2 (92 percent), while the agreement rate for non-Hispanics is higher in Method 2 (99 percent) compared to Method 1 (97 percent). For this paper, we chose to further evaluate Method 1 using multinomial regressions since the agreement rate is higher for Hispanics compared to Method 2.

4.1.2. Race

We explored six different methods of assigning a single race response, and they are described above. Figure 2 shows how we applied the business rules described in each method to assign one race response from administrative records. If there is no discrepancy in an individual's race responses across files, then that race was assigned to the administrative records composite. This is reflected in the first two rows of Fig. 2. As shown in Table 1, there were 274 million (78 percent) records in the composite with no discrepancy across race responses. About 91 million (26 percent) had only one source of race data, and 183 million (52 percent) had the same race response across multiple sources.

Approximately 9 million (3 percent) of individuals in the composite had different race responses. If responses across source files are discrepant, race was assigned in the following manner:

Methods 1 and 2 prioritized smaller race groups over larger ones, since smaller race groups – specifically, the AIAN, Asian, NHPI, SOR, and multiracial groups – experience lower race response agreement and greater coverage issues than Whites and Blacks [6,7,26]. A single race was assigned with preference given to smaller race groups according to their share of the total 2010 Census population distribution. As NHPI alone is the smallest of the seven race categories, it was selected first if it was in any of the source files, as demonstrated in rows 3 and 4 of the figure. Then race was selected in the following order – AIAN alone, Two or More Races, Asian alone, SOR alone, Black alone, and then White alone. Method 2 is similar to Method 1, but preference was given to Two or More Races first, followed by NHPI alone, AIAN alone, Asian alone, SOR alone, Black alone, and then White alone. This is demonstrated in row 3, in the "Method 2" column of Fig. 2. We chose to prioritize Two or More Races first in this method because this group experiences lower agreement and coverage rates in ARTPD relative to other groups [6].

For Methods 3 and 4 we prioritized the race response that was most frequently reported across ARTPD. For

Table 2
Percentage of ARTPD Hispanic origin data matched to the 2010 Census

Hispanic origin responses	Method 1		Method 2	
	Number	Percent	Number	Percent
Hispanic	30,253,046	93.7	29,713,244	92.0
Non-Hispanic	191,087,892	96.8	195,097,141	98.8

Source: 2010 Census and ARTPD. Note: ARTPD = administrative records and third party data.

Table 3
Percentage of ARTPD race data matched to the 2010 Census

Race responses	Method 1		Method 2		Method 3		Method 4		Method 5		Method 6	
	Number	Percent	Number	Percent	Number	Percent	Number	Percent	Number	Percent	Number	Percent
White alone	161,093,312	96.5	161,093,312	96.5	161,748,097	96.9	161,748,097	96.9	161,817,315	96.9	161,745,421	96.9
Black alone	26,660,073	96.4	26,660,073	96.4	26,776,690	96.8	26,776,690	96.8	26,706,261	96.6	26,702,739	96.6
AIAN alone	1,494,452	74.3	1,414,241	70.3	1,430,924	71.1	1,385,324	68.9	1,357,520	67.5	1,466,302	72.9
Asian alone	6,598,137	90.8	6,598,137	90.8	6,263,855	86.2	6,263,855	86.2	6,624,789	91.1	6,624,401	91.1
NHPI alone	146,565	57.6	143,683	56.5	138,968	54.6	136,442	53.6	144,115	56.7	144,049	56.6
SOR alone	2,656,812	49.4	2,656,812	49.4	2,592,965	48.2	2,592,965	48.2	2,627,712	48.8	2,626,585	48.8
Two or More Races	1,367,434	28.6	1,450,139	30.4	1,272,013	26.6	1,329,544	27.8	1,440,223	30.2	1,385,900	29.0

Source: 2010 Census and ARTPD. Note: ARTPD = administrative records and third party data.

Method 3, this is demonstrated in rows 3, 5, and 6 of the column labeled "Method 3" of Fig. 2. If there was no most frequent race, then race was assigned with preference given to smaller race groups using the same order as in Method 1. This is demonstrated in row 4 of the figure. For Method 4, if there was no most frequent race, then race was assigned with preference given to the smaller race groups using the same order as in Method 2. This is demonstrated in row 7 of the figure.

In Methods 5 and 6, we evaluated whether dataset order impacted agreement rates. Since previous census records are most similar in format and design to the 2010 Census, we assigned a race response to the ARTPD composite first from this data source. This is shown in rows 3, 5, 6 and 7 of the "Method 5" column. If a response was missing in previous census records, then a race response was assigned according to Method 1. This is reflected in row 4. In Method 6, we evaluated whether using IHS before other datasets would increase agreement response rates for American Indians and Alaska Natives. This is illustrated in rows 4, 5, and 6. Otherwise, a race response was assigned as described in Method 5. Rows 3 and 7 of the "Method 6" column provide examples of this.

As with the Hispanic origin assignment, if there are no race responses in administrative records, then race is set to missing in the composite file regardless of the method used. This is illustrated in the last row of Fig. 2. There were 69 million (20 percent) individuals in the composite with missing race information (see Table 1).

The match rates for each race method are shown in Table 3. Method 1 generally looks like the best method. Method 1 has the highest agreement rates for AIAN and NHPI and very similar agreement rates for single race Whites, Blacks, and those who report SOR alone compared to other methods. The agreement rate for Asians is very similar to some of the other methods; Methods 3 and 4 have much lower agreement rates for Asians compared to the other methods (86 percent compared to 91 percent, respectively). The agreement rate for Two or More Races is also quite similar across the methods, but Methods 2 and 5 have slightly higher agreement rates (30 percent) compared to Method 1 (29 percent).

4.2. Regression results

Next, using multinomial regressions, we discuss factors associated with matching, non-matching, and missing Hispanic origin (model 1) and race (model 2) responses between the 2010 Census and ARTPD. Table 4 presents the distribution of the dependent variables. Overall, about 87 percent of individuals have matching Hispanic origin responses, 3 percent have non-matching or different Hispanic origin responses, and 10 percent have missing Hispanic origin data in ARTPD. A smaller percentage of respondents have matching race responses at 78 percent. Five percent of respondents have different race responses between the two data sources and 17 percent are missing race data in administrative records.

Model 1 shows that Hispanics are more likely to have non-matching Hispanic origin responses but less

Table 4
Distribution of dependent variables used in multinomial logistic regression models

	Hispanic origin		Race	
	Number	Percent	Number	Percent
Total	251,320,952	100.0	253,905,696	100.0
Matching	217,393,894	86.5	198,805,047	78.3
Non-Matching	8,068,227	3.2	11,707,695	4.6
Missing ARTPD Data	25,858,831	10.3	43,392,954	17.1

Source: 2010 Census and ARTPD. Note: ARTPD = administrative records and third party data.

Table 5
Multinomial logistic regression results, odds ratios

	Model 1: Hispanic origin (Matching Hispanic origin is the reference)		Model 2: Race (Matching race is the reference)	
Variables	Non-matching responses	Missing in ARTPD	Non-matching responses	Missing in ARTPD
Ethnicity/Race in Census (Non-Hispanic White alone omitted)				
Hispanic	1.58***	0.71***	35.95***	9.24***
Non-Hispanic Black alone	0.67***	0.88***	1.63***	0.71***
Non-Hispanic American Indian or Alaska Native alone	2.20***	0.89***	13.67***	0.41***
Non-Hispanic Asian alone	2.51***	2.98***	6.29***	10.88***
Non-Hispanic Native Hawaiian or Pacific Islander alone	3.60***	2.01***	35.95***	9.24***
Age (18–44 years old omitted)				
0–17 years old	1.24***	12.66***	2.19***	16.85***
45–64 years old	0.96***	0.40***	0.95***	0.48***
65 years and older	0.83***	0.24***	0.86***	0.25***
Gender (Female omitted)				
Male	0.74***	1.07***	1.05***	1.00***
Household Tenure (Owner omitted)				
Renter	1.08***	1.04***	1.16***	1.63***
No rent paid	1.09***	1.08***	1.09***	1.48***
Household Type (Married couple family omitted)				
Single father family	1.12***	0.61***	1.07***	0.96***
Single mother family	1.18***	0.43***	1.02***	0.58***
Other household type	1.06***	0.97***	1.13***	1.02***
Household Size (1 person omitted)				
2 persons	1.06***	1.04***	1.03***	1.42***
3 persons	1.13***	1.32***	1.10***	2.22***
4 persons	1.11***	1.30***	1.10***	2.42***
5 or more persons	1.07***	1.07***	1.18***	2.64***
Census Mode (Mailout/Mailback omitted)				
Nonresponse follow-up	1.23***	1.13***	1.48***	1.27***
Other mode	1.06***	1.02***	1.06***	1.23***
Region (West omitted)				
Midwest	0.63***	0.84***	0.77***	1.03***
South	0.76***	0.99***	0.77***	1.07***
Northeast	0.97***	0.94***	0.94***	1.09***
Address Type (Urban omitted)				
Rural	0.76***	0.93***	0.85***	0.93***
Percent non-Hispanic White in tract	0.99***	1.00***	1.00***	1.00***
Median household income in tract (log)	1.39***	2.06***	1.00	1.19***
Unweighted N	251,320,952		253,905,696	

***$p < 0.001$; Source: 2010 Census and ARTPD; Note: ARTPD = administrative records and third party data.

likely to have missing Hispanic origin data in ARTPD compared to non-Hispanic Whites (Table 5). In model 2, we see that Hispanics are 43 times more likely to have non-matching race responses in ARTPD than non-Hispanic Whites. Also, the odds of having missing race responses are about 11 times larger for Hispanics than non-Hispanic Whites. One factor contributing to these results is that many Hispanics view their race as "Hispanic" and do not identify with OMB's standard race groups.

With the exception of non-Hispanic Black individuals in the Hispanic origin model, non-Hispanic mi-

norities are significantly more likely to have different Hispanic origin and race responses in ARTPD than in the Census compared to non-Hispanic Whites. In fact, the odds of having non-matching race responses are 14 and 36 times larger for non-Hispanic AIAN and non-Hispanic NHPI individuals, respectively, relative to non-Hispanic Whites. This is consistent with previous research that shows that response change is more common in AIAN and NHPI groups [10–12]. The coefficients in both models show that race and Hispanic origin data in ARTPD is more likely to be missing for non-Hispanic Asians and non-Hispanic NHPIs than for non-Hispanic Whites. These results are supported by previous research which finds that coverage in administrative records is lower for minorities compared to non-Hispanic Whites [6,7,26].

There are differences in matching race and Hispanic origin by age and gender. Individuals aged 45 and older are more likely to have the same race and Hispanic origin responses in the Census and ARTPD, however, individuals aged 17 and younger are less likely to have matching responses than those aged 18 to 44 years. This pattern is consistent with previous studies measuring agreement of race and Hispanic origin data in surveys and administrative records [19,20]. Younger individuals are 17 and 13 times more likely to have missing race and Hispanic origin administrative records data, respectively. These results are consistent with findings from previous research that coverage in administrative records is lower for younger age groups compared to older age groups [5,6]. Males are less likely to have non-matching Hispanic origin responses. This contradicts earlier research that finds that males have a lower likelihood of having consistent ethnicity responses than females [18,19]. Consistent with prior research using administrative records data, males are more likely to have non-matching race responses than females [18,19]. In addition, males are more likely to have missing Hispanic origin data.

Compared to individuals who responded to the 2010 Census by mail, individuals who responded to the Census in the nonresponse follow-up operation or other modes are more likely to have non-matching race and Hispanic origin responses in ARTPD.[4] The presence of an enumerator in the 2010 Census nonresponse follow-up operation may affect a respondent's response to the Hispanic origin and race questions. In a study of Hispanic origin and race response change between Census 2000 and the 2010 Census, Liebler et al. [10] found that response change is common among people who responded to one or both censuses using a response mode other than mail. Responding to the Census by non-mail modes is associated with a greater likelihood of missing race and Hispanic origin data in ARTPD.

Race and Hispanic origin response agreement between the 2010 Census and ARTPD varies by geography. Consistent with findings from earlier research, residents living in the West are more likely to have non-matching Hispanic origin and race responses than those residing in the Midwest, South, or Northeast [10]. Individuals living in regions other than the West are less likely to have missing Hispanic origin responses but more likely to have missing race data. In addition, living in rural areas is associated with matching responses for both race and Hispanic origin. Minority groups are more likely to have non-matching race responses compared to Whites [10] and are also more likely to live in urban areas [27], which may in part be why we observe more matching responses in rural areas.

In terms of household characteristics, renters, households headed by single parents or that have other household compositions, and individuals living in households with two or more people are more likely to have different Hispanic origin and race responses compared to homeowners, households headed by married couples, and individuals who live alone. Households headed by single parents are less likely to have missing race and Hispanic origin data than households headed by married couples. Renters and individuals living in households with two or more people are more likely to have missing race and Hispanic origin data in ARTPD than homeowners and single person households.

As the median household income in the tract of residence increases, the odds of having a non-matching or missing Hispanic origin response in ARTPD increase. Individuals living in more affluent neighborhoods are more likely to have missing race data than those in neighborhoods with lower median household incomes. These findings are consistent with the work of Fernandez et al. [18].

5. Conclusion

In this research, we explored different methods for assigning one race and Hispanic origin response when

[4] Our analyses include census information collected from a neighbor or other respondent outside of the household (i.e., proxy reports). We also ran our regressions excluding proxy reports and generally found that the patterns were similar.

responses are discrepant across ARTPD sources. We evaluated which methods resulted in the highest level of agreement with the 2010 Census and find that Hispanic Method 1 and race Method 1 resulted in the highest match rates for Hispanics and smaller race groups, respectively. For Hispanic origin, the match rate was 94 percent for Hispanics and 97 percent for non-Hispanics. The most successful method in assigning race resulted in match rates ranging from 29 percent for Two or More Races to 96 percent for White alone and Black alone. Match rates were higher for single race Whites, Blacks, and Asians and lower rates for single race AIANs, NHPIs, those who report SOR, and Two or More Races.

We also described the characteristics of individuals whose Hispanic origin and race responses in ARTPD do not match 2010 Census data or are missing. We find that many demographic, household, and contextual variables are associated with non-matching and missing race and Hispanic origin responses. The magnitude of race, ethnicity, age, and household size odds ratios are notable.

We find that minorities, especially Hispanics, are more likely to have non-matching Hispanic origin and race responses in ARTPD compared to the 2010 Census. These results are consistent with those found in other studies on racial and ethnic fluidity. Hispanics are less likely to have missing Hispanic origin data but more likely to have missing race data in ARTPD. Hispanics' higher likelihood of missing race data in ARTPD relative to non-Hispanic Whites may be in part due to not identifying with the response options offered. We also find that non-Hispanic Asian and NHPI individuals are more likely to have missing race and Hispanic origin data in ARTPD.

Consistent with previous research that children are not covered as well as adults in ARTPD [6,7], individuals ages 0–17 are more likely to have missing race and Hispanic origin responses. Household size is also strongly associated with missing race data, where larger households tend to have missing race data compared to smaller households. This may also be related to coverage of children issues in ARTPD.

Our findings suggest that using ARTPD when race and Hispanic origin responses are missing is a promising approach. The quality of ARTPD is high for the White alone, Black alone, and Asian alone populations, and these data can assist in assigning responses when data are missing. Our results concur with those found during Census 2000 research and show that ARTPD race and Hispanic origin responses for minority groups had lower levels of agreement with 2010 Census data relative to non-Hispanic Whites. Although the quality of race and Hispanic origin administrative data appears to be lower for some minority groups, our results are consistent with earlier research on racial and ethnic fluidity. Further research is needed to investigate scenarios in which ARTPD should be used for imputation given ARTPD could impact Hispanic origin and race distributions [3].

As the Census Bureau acquires more administrative records and third party data sources, we will evaluate the agreement of Hispanic origin and race response data relative to census data in order to assess the use of these sources in the ARTPD race and Hispanic origin composite. We are currently in the process of acquiring SNAP and Women, Infants, and Children program data from states and will evaluate these data for the ARTPD composite. We will also explore additional methods to assign race and Hispanic origin from ARTPD.

By developing methods to assign one Hispanic origin and race response when these responses are discrepant across administrative records sources, our research can inform imputation strategies to address race and ethnicity item nonresponse in census surveys and the 2020 Census. Our research can also inform other Census Bureau programs including the Population Estimates Program where national population estimates by race and ethnicity are developed using administrative records. By contributing to a better understanding of the factors associated with non-matching and missing race and Hispanic origin data in ARTPD, our analysis will inform the application of ARTPD race and Hispanic origin data to Census operations and research as well as research on race and Hispanic origin reporting, measurement, and fluidity.

Acknowledgments

The authors would like to thank Leticia E. Fernandez and Amy O'Hara for their helpful comments and suggestions.

References

[1] Farber J, Wagner D, Resnick D. Using administrative records for imputation in the decennial census. In JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association, 2005; 3008-3015.

[2] Rothhaas C, Lestina F, Hill JM. 2010 Decennial Census: Item nonresponse and imputation assessment report. 2010 Census Planning Memoranda Series No. 173, 2012.

[3] Rastogi S, Fernandez L, Noon J, Zapata E, Bhaskar R. Exploring administrative records use for race and Hispanic origin item non-response, Center for Administrative Records Research and Applications Working Paper No. 2014-16, Washington, DC: U.S. Census Bureau, 2014.
[4] Bye B, Judson D. Results from the Administrative Records Experiment in 2000. Census 2000; Synthesis Report No. 16. Washington, DC: U.S. Census Bureau, 2004.
[5] Farber J, Leggieri C. Building and validating a national administrative records database for the United States. Paper presented at New Zealand Conference on Database Integration, 2002.
[6] Rastogi S, O'Hara A. 2010 Census Match Study. 2010 Census Planning Memoranda Series No. 247. Washington, DC: U.S. Census Bureau, 2012.
[7] Bhaskar R, Luque A, Rastogi S, Noon J. Coverage and agreement of administrative records and 2010 American Community Survey demographic data. Center for Administrative Records Research and Applications Working Paper No. 2014-14, Washington, DC: U.S. Census Bureau, 2014.
[8] Doyle JM, Kao G. Are racial identities of multiracials stable? Changing self-identification among single and multiple race individuals. *Soc Psychol Q.* 2007; 70(4): 405-423.
[9] Harris D, Sim JJ. Who is multiracial? Assessing the complexity of lived race. *Am Sociol Rev.* 2002; 67(4): 614-627.
[10] Liebler CA, Porter SR, Fernandez LE, Noon J, Ennis SR. America's churning races: race and ethnicity response changes between Census 2000 and the 2010 Census. *Demography.* 2017; 54(1): 259-284.
[11] Bentley M, Mattingly T, Hough C, Bennett C. Census Quality Survey to evaluate responses to the Census 2000 question on race: an introduction to the data. Census 2000 Evaluation B.3. Washington, DC: U.S. Census Bureau, 2003.
[12] del Pinal J, Schmidley D. Matched race and Hispanic origin responses from Census 2000 and Current Population Survey February to May 2000. Population Division Working Paper No. 79; Washington, DC: U.S. Census Bureau, 2005.
[13] Dusch G, Meier F. 2010 Census Content Reinterview Survey evaluation report. 2010 Census Program for Evaluations and Experiments. Washington, DC: U.S. Census Bureau, 2012.
[14] Singer P, Ennis SR. Census 2000 Content Reinterview Survey: accuracy of data for selected population and housing characteristics as measured by reinterview. Census 2000 Evaluation B.5. Washington, DC: U.S. Census Bureau, 2003.
[15] U.S. Census Bureau. Content Reinterview Survey: accuracy of data for selected population and housing characteristics as measured by reinterview. 1990; Census of Population and Housing Evaluation and Research Reports. Washington, DC: U.S. Census Bureau, 1993.
[16] Brown JS, Hitlin S, Elder GH, Jr. The greater complexity of lived race: an extension of Harris and Sim. *Soc Sci Q.* 2006; 87(2): 411-431.
[17] Campbell ME, Rogalin CL. Categorical Imperatives: the interaction of Latino and racial identification. *Soc Sci Q.* 2006; 87(5): 1030-1052.
[18] Fernandez L, Rastogi S, Ennis SR, Noon J. Evaluating race and Hispanic origin responses of Medicaid participants using census data. Center for Administrative Records Research and Applications Working Paper No. 2015-01; Washington, DC: U.S. Census Bureau, 2015.
[19] McAlpine DD, Beebe TJ, Davern M, Call KT. Agreement between self-reported and administrative race and ethnicity data among Medicaid enrollees in Minnesota. *Health Serv Res.* 2007; 42(6): 2373-2388.
[20] Gomez SL, Kelsey JL, Glaser SL, Lee MM, Sidney S. Inconsistencies between self-reported ethnicity and ethnicity recorded in a health maintenance organization. *Ann Epidemiol.* 2005; 15(1): 71-79.
[21] Kressin NR, Chang B, Hendricks A, Kazis LE. Agreement between administrative data and patients' self-reports of race/ethnicity. *Am J Public Health.* 2003; 93(10): 1734-1739.
[22] Office of Management and Budget. Revisions to the standards for the classification of federal data on race and ethnicity. 1997.
[23] Fellegi IP, Sunter AB. A theory for record linkage. *J Am Stat Assoc.* 1969; 64: 1183-1210.
[24] Wagner D, Layne M. The Person Identification Validation System (PVS): applying the Center for Administrative Records Research & Applications record linkage software. Center for Administrative Records Research and Applications Working Paper No. 2014-01. Washington, DC: U.S. Census Bureau, 2014.
[25] Bond B, Brown JD, Luque A, O'Hara A. The nature of bias when studying only linkable person records: evidence from the American Community Survey. Center for Administrative Records Research and Applications Working Paper No. 2014-08. Washington, DC: U.S. Census Bureau, 2014.
[26] Luque A, Bhaskar R. 2010 American Community Survey Match Study. Center for Administrative Records Research and Applications Working Paper No. 2014-03. Washington, DC: U.S. Census Bureau, 2014.
[27] U.S. Census Bureau. American FactFinder [Internet]. 2010 Census Summary File 1; Table P5 [cited 16 January 2015]. Available from: http//factfinder2.census.gov.