DEALING WITH INCONSISTENT RACE DATA IN THE PRODUCTION OF POPULATION ESTIMATES: IMPROVEMENTS AT THE SUBNATIONAL LEVEL

by Amy Symens Smith
Population Division
U.S. Census Bureau

For presentation at the annual meeting of the Population Association of America, April 1-4, 2004, Boston, MA

ABSTRACT

Changes in the U.S. racial categories implemented in Census 2000, including multiple race reporting, have made it necessary to modify the Census Bureau's methodology to produce postcensal population estimates by race. A race modeling technique has been implemented that ensures consistency across racial categories while the various administrative data systems transition to the new race standards.

The current model for estimating births is based on Census 2000 data at the national level for the population under age one and their parents. However, Census 2000 data indicate important regional and state variations in race reporting. The purpose of this paper is to explore modeling at lower levels of geography with the aim to improve subnational population estimates. This research outlines a statistical model to incorporate such differences and provides a first look at descriptive statistics.

BACKGROUND

The Office of Management and Budget (OMB) in 1997 issued revised standards for collecting, tabulating and presenting data on race and Hispanic origin.[1] The race categories were expanded to include: White; Black or African American; American Indian or Alaska Native; Asian; and, Native Hawaiian and Other Pacific Islander. Additionally, respondents were given the option, for the first time, to mark more than one race. The new race standards were used in Census 2000 and other Federal programs were mandated to adopt the standards as soon as possible, but no later than January 1, 2003.

Starting with the 2000 population estimates and continuing to the present, it has been necessary to reexamine the methodology used to estimate the population by race. In particular, we had to reexamine the methodology used to assign race to births. Initially, race was modeled by distributing births according to the race/Hispanic origin distribution of the Census 2000 age zero population. At the time, analysis of the Census 2000 data was underway, but little information was known about the multiple race population. Distributing post-April 1, 2000 births according to the Census distribution was based on a simple method that ensured a reasonable race/Hispanic origin distribution for the estimated age zero population. However, the obvious shortcoming was that the population born after Census Day experienced no real change from factors such as differential fertility rates by race, immigration, or respondents' changing ideas about their race.

The model currently in use is more sophisticated and incorporates knowledge from Census 2000 about multiple race reporting in the United States. Previous research indicated that it was no longer sufficient to determine whether the child's race followed

[1] Federal Register Notice 10/30/97 Vol. 62 No. 210 Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity.

either the mother's race (mother-rule) or the father's race (father-rule).[2] With the option to indicate multiple races the child's race could now be reported as a combination of both the parents' races, thus, the multiple race-rule was introduced.

The current modeling technique is based on within household race reporting between children under age one and their parents, as reported in Census 2000. The model and its development is summarized below; see earlier research for a more detailed description.[3]

1. The Census 2000 relationship question was used to select children under age one who were the natural-born sons and daughters of the householder.[4]
2. The subsetted Census 2000 data was cross-tabulated to produce a matrix of the child's race/Hispanic origin by the father's race/Hispanic origin by the mother's race/Hispanic origin.
3. Using birth certificate data, births were cross-tabulated by father's race/Hispanic origin and mother's race/Hispanic origin.
4. Births cross-tabulated by parent's race/Hispanic origin (step #3) were assigned a race/Hispanic origin based on the race/Hispanic origin distribution of the age zero population (step #2) for the matching parent group.

The model was evaluated in the production of the Census Bureau's recent population estimates by comparing the July 1, 2001 estimated age zero population (modeled births) and the estimated age one population (approximately age zero in Census 2000). In brief, the results showed that overall differences in the race distribution were small; however; differences were larger when looking at the full race by Hispanic origin distribution. Modeling resulted in smaller percentages of non-Hispanic White alone births and larger percentages of non-Hispanic Black alone and Hispanic White alone births.

Research is underway to learn more about these differences and to improve the technique. One consideration is to limit the modeled race of birth to be consistent with one or both parents on race and/or Hispanic origin. Currently the model does not limit race so, for instance, a White mom and a Black dad may have an Asian child if this scenario had been reported in the Census. However, for most groups more than 90 percent of children have a race/Hispanic origin that is consistent with one or both parents. Where differences do exist, they are the largest for the Hispanic population.

Another possible explanation for the differences between modeled births and the age one population is differential race reporting by parents in the Census and on birth certificates.

---

[2] I Wanna Be Like Mike Tiger Woods! Exploratory Analysis of Race Reporting for Children in Interracial Households in Census 2000. Amy Symens Smith and Nicholas A. Jones. 2001. Paper presented Southern Demographic Association meetings October 11-14, 2001. Miami, FL.

[3] Dealing with the Changing U.S. Racial Definitions: Producing Population Estimates Using Data with Limited Race Detail. Amy Symens Smith and Nicholas A. Jones. 2003. Paper presented at the Population Association of America meetings May 1-4, 2003. Minneapolis, MN.

[4] The "natural-born son/daughter" of the householder (person1) is the son or daughter of the householder by birth regardless of age of child.

Perhaps race/Hispanic reporting in the Census and on birth certificates is different. There is evidence that this is particularly true for the American Indian population.[5]

Finally, differences for some race groups, and for the Hispanic population, may indicate Census undercount. Previous censuses indicate that undercount varies by race, and is higher for minority groups. On the other hand, the vital statistics system is considered to be nearly 100 percent complete.

The National Center for Health Statistics (NCHS) and the Census Bureau have worked closely this decade to solve the problems associated with using two different race classification systems. The next section describes recent research by NCHS to "bridge" to consistent race groups.

NCHS RESEARCH

During the transition to the new race standards, NCHS needed a "bridging" technique to help construct vital rates published annually. Starting with the 2000 data year, the vital rates numerators (vital statistics) and denominators (postcensal estimates) had incomparable race data. Early in the decade the Census Bureau assisted by producing 1990-based July 1, 2000 and July 1, 2001 population estimates consistent with the old race standards.

However, this was not a long-term solution because 1990-based estimates were not consistent with the Census 2000 enumeration. Thus, NCHS devised a bridging technique that relied on four years of National Health Interview Survey (NHIS) data.[6] The NHIS is an ideal data source because since 1982 it has allowed respondents to choose more than one race, and then followed-up by asking multiple race respondents for their primary or "best" single race.

Logistic regression models were fit to the NHIS data that included both demographic and contextual covariates. All models included age in single years, sex, and Hispanic origin (Hispanic or not Hispanic). County-level contextual variables included county of residence, region, level of urbanization and percent of county population that reported more than one race. County-level single race population percents were used in the appropriate models. Regression coefficients predicted the probability of selecting a specified category as one's primary race. Ingram et al. concluded "that the NHIS regression method is a better predictor of primary race than other methods because it incorporates covariate information and thus adjusts for variations across counties in the distribution of age, gender, Hispanic origin, and multiple-race groups" p 13.

---

[5] See Race and Ethnicity Classification Consistency between the Census Bureau and the National Center for Health Statistics. Larry D. Sink 1997. Population Division Working Paper No. 17.

[6] See Methodology for Bridging Race in the Modified Race Data Summary File for Census 2000 and Subsequent Population Estimates. Deborah D. Ingram, James A. Weed, Jennifer Parker, Nathaniel Schenker and Jennifer Madans. 2003. National Center for Health Statistics.

Borrowing from NCHS' research, the Census Bureau's race modeling methodology can benefit by implementing statistical modeling to assign race of birth. Previous research examining Census 2000 data shows variations in race reporting consistent with the demographic and contextual variables introduced by NCHS. The next section describes this research.

RELATED RESEARCH

Previous research has examined race reporting for children in households, comparing the child's race(s) with the parents' race(s). This research revisited the traditional race reporting paradigms to determine if the child's race followed the mother-rule or the father-rule. Additionally, this research introduced a new paradigm, the multiple race-rule, to determine if the child's race was reported as the combination of the parents' race(s). White ***and*** Asian children were the most likely to follow the multiple race-rule with 93 percent of children doing so. Conversely, White ***and*** Black children where the least likely to have multiple race responses, following from their White and Black parentage.[7]

This research looked at multiple race reporting by geography. The race reporting for children in White ***and*** American Indian and Alaska Native (AIAN) families showed regional differences. The reporting of White alone was 36 percent overall, and much more prevalent in the Northeast (42 percent) and Midwest (40 percent). The reporting of AIAN alone (35 percent overall) showed the highest levels in the South (33 percent) but was much lower in the Northeast (18 percent). On the other hand, the reporting of White ***and*** AIAN was similar across regions, ranging from a low of 32 percent in the South to a high of 38 percent in the West and Northeast.

The reporting of race for children living in White ***and*** Asian families was somewhat similar across regions. In each region, the reporting of single races was much lower than the reporting of White ***and*** Asian. Overall, the reporting of White alone was 28 percent, with a low of 23 percent in the West and a high of 36 percent in the South. The reporting of Asian alone was 13 percent overall and did not vary much by region. However, the reporting of White ***and*** Asian was very high overall (54 percent) ranging from 45 percent in the South to a high of 60 percent in the West.

Finally, we found that the reporting of race for children living in White ***and*** Black families was also somewhat similar across the four regions. The reporting of White alone was 18 percent overall with a low of 15 percent in the Midwest and a high of 23 percent in the Northeast. The reporting of Black alone was 28 percent overall and highest in the

[7] I Wanna Be Like Mike Tiger Woods! Exploratory Analysis of Race Reporting for Children in Interracial Households in Census 2000. Amy Symens Smith and Nicholas A. Jones. 2001. Presented at the Southern Demographic Association meetings, October 11-14, 2001. Miami, FL. Also see Who is 'Multiracial?" Exploring the Complexities and Challenges Associated with Identifying "the" Multiracial Population in Census 2000. Nicholas A. Jones and Amy Symens Smith. 2002. Presented at the Population Association of American meetings, May 9-11, 2002. Atlanta, GA.

South (31 percent), with the other three regions at 27 percent. The reporting of White ***and*** Black was very high overall (44 percent), and highest in the West (49 percent) and Midwest (48 percent).

Incorporating factors that allow for race reporting variations such as these will improve the race modeling technique. The next section outlines the improved race modeling technique and additional demographic and contextual variables.

RACE MODELING IMPROVEMENTS

Logistic regression models can be fit to the subsetted Census 2000 data previously used to capture race reporting relationships in households between parents and their children. Multinomial logistic regression models would be constructed for each combination of father's and mother's race(s) with the model predicting the child's race. The resulting regression coefficients would predict the probability of the child's race following the mother-rule, father-rule or the multiple race-rule. The probabilities would be used to assign a race to the birth component data.

To illustrate, separate models can be constructed for each combination of father and mothers race, or combined for combinations with similar outcomes. One model may be constructed for the interracial combination of Black dad and White mom. The model would predict whether the child's race was reported as White alone (mother rule), Black alone (Father rule) or White ***and*** Black (multiple race rule).

The benefit in introducing statistical modeling is that covariates such as demographic, geographic and contextual variables can be introduced. To make improvements at the subnational level, these indicators can be tabulated at the state and county levels. Variables must be present in both the Census 2000 subsetted data and in the birth certificate data. Figure 1 displays the model used to predict race reporting for children in Black/White interracial families. Independent variables include: sex of child (male/female) and combined age of parents (mother's age + father's age). Contextual variables include: state (state fips code) and percent race group in the state (race group(s) vary according to the model).

The scope of this paper only permits for the presentation of descriptive statistics, which are presented below. The full multinomial logistic regression models will be constructed in future research.

DESCRIPTIVE STATISTICS

Here we focus on one of the most common interracial families in the United States, those with a Black alone dad and a White alone mom. Using Census 2000 data to identify natural-born sons and daughters in two-parent families resulted in 35,452 under age one children. Table 1 shows the White *alone*, Black *alone* and White ***and*** Black multiple

race reporting for these children. Nearly half of all children (48 percent) were reported as White ***and*** Black. The second most common race reported was Black alone (35 percent), in this case following the father-rule and/or the minority race-rule. An additional 16 percent of children were reported as White alone.[8]

The current model assigns race to babies based on distributions similar to the one illustrated above. That is, for babies with a Black *alone* dad and a White *alone* mom, 48 percent are assigned White ***and*** Black race, 35 percent are assigned Black ***alone*** race, and 16 percent are assigned White ***alone*** race. At present this assignment of race does not control for any demographic or contextual variables, which may influence race reporting.

First looking at demographic characteristics, Table 2 shows little variation by sex when reporting the race of child in Census 2000. Both boys and girls (48 percent and 49 percent) are most likely to be reported as White ***and*** Black multiple race. The largest difference by sex is in the reporting of Black *alone* with a larger percentage of boys (36 percent) then girls (34 percent) with this race. This suggests a greater preference for boys than girls to follow the father-rule, or in this case the minority race-rule.

Figure 2 looks at the combined age of parents[9] and the reporting of race for the child. At nearly all combined ages the preference is for reporting child's race as White ***and*** Black. Starting at combined age 85, and for several combined ages above that, the child's race is often reported as Black *alone*. However, at these higher combined ages there are fewer children, which may impact the quality of the data.

Turning to the geographic variables, Table 3 shows that it is important to expand the model to take into consideration state variations. For the majority of states White ***and*** Black is the preferred race. However, for six states – South Dakota, New Hampshire, New York, Rhode Island, Louisiana, Texas and New Mexico- this is not the case. In all but one of these states, the preference is for Black *alone*. Census 2000 data shows that Texas, Louisiana and New York are three of the ten states with the largest Black *alone* or in combination populations.[10] At the other extreme, Census 2000 data shows that South Dakota and New Hampshire are two of the 13 states with less than 3 percent of the total state population reporting Black *alone* or in combination.

Finally, Table 4 shows the concentration of the White ***and*** Black population in states and the reporting of race for the child. A concentration variable was created for each state by calculating the White ***and*** Black population as a percent of the total population. Concentration of the White ***and*** Black population ranged from 0.47 percent to 0.12 percent; quintiles were used to construct the table. In states with the highest concentration of White ***and*** Black population, as well as states with the lowest

[8] Approximately a half of a percent of children with a Black dad and a White mom had a race other than White alone, Black alone or White and Black. In the new modeling recommendation we are suggesting that race a of child only be modeled a single or multiple race that is consistent with the parents' race(s).
[9] A combined age of parent variables is created by adding the mother and father's age. Values range from 31 years to 127 years.
[10] See the Census 2000 Brief: The Black Population: 2000. Jesse McKinnon. U.S. Census Bureau.

concentration of White ***and*** Black population, children are most likely to be reported as White ***and*** Black. However, for states with the second lowest level of concentration (ranging from 0.28 to 0.20 percent) Black *alone* is slightly more preferred than White ***and*** Black.

DISCUSSION

The purpose of this paper was to recommend an improvement in the current race modeling technique used to assign race to births. It is quite possible that for the next few years, while administrative data systems transition to the new race standards, data will be tabulated using both the new and old race standards. For this reason, it has been necessary to explore race modeling procedures to produce population estimates.

The race modeling methods that have been developed so far have been implemented at the national level, applying race reporting relationships apparent in Census 2000. This research has set a strong foundation for race modeling, however, it is now time to build-upon this foundation by introducing enhancements to the model which will improve population estimates below the national level.

This paper introduced a technique used by the National Center for Health Statistics, which can be adapted for use at the Census Bureau. NCHS used statistical modeling to determine the probabilities of choosing a particular single race outcome for respondents that first provided a multiple race response. A similar technique can be implemented for use in race modeling. Models can be fit to the Census 2000 data currently used to capture race reporting relationships between parents and children. The outcomes of the models would be coefficients that can be used to assign races to births.

The demographic, geographic and contextual variables introduced above are a first look at variables which may prove to be statistically significant predictors in the statistical models. Variations in reporting are apparent when considering state indicator as well as concentration of the particular race groups. On the other hand, variables such as gender and combined age of parents show little variation

Statistical modeling also provides the option to test for statistically significant differences between models. It is anticipated that a model will be constructed for each combination of parents' races. However, this may be cumbersome considering that to describe interracial Black/White families alone requires eight different models. By using statistical methods to determine if there are significant differences between these eight models, collapsing can be used, if appropriate.

It is clear that modeling race is a challenging and time consuming effort. A full transition to consistent race categories by all administrative systems will clearly ease the production of postcensal population estimates. The availability of birth certificate data where multiple race reporting is an option will provide a wealth of information on the way that parents and their children report race and conceptualize race within family units. Use of

such information in producing population estimates, at the national level and below, will improve the accuracy of the current population estimates

Figure 1. Model of race reporting for children under age one with a Black *alone* dad and a White *alone* mom

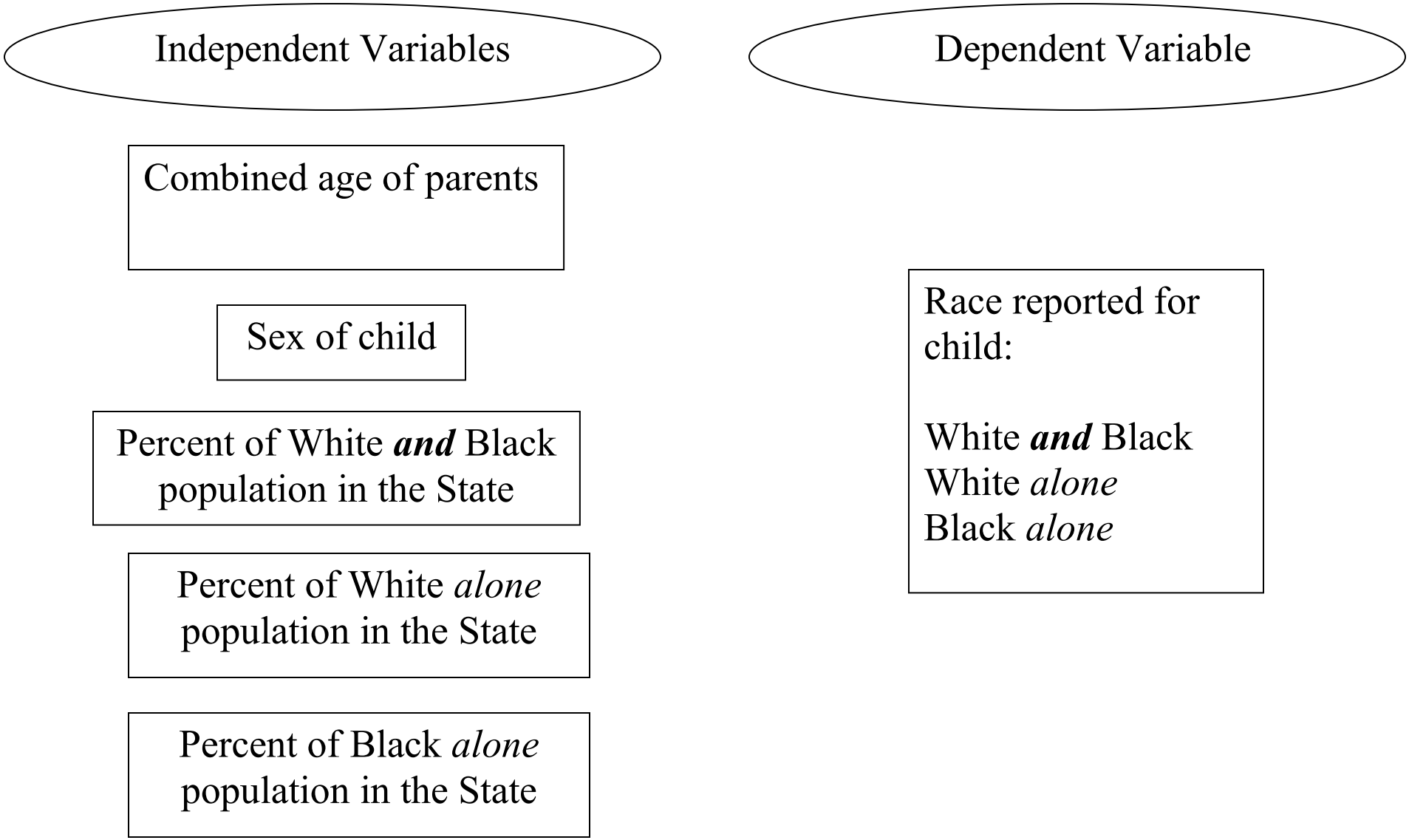

Table 1. Race reporting for children under age one in families with a Black *alone* dad and a White *alone* mom

Race of Child	Number	Percent
Total	35,452	100.00
White *alone*	5,708	16.10
Black *alone*	12,451	35.12
White ***and*** Black	17,124	48.30

Numbers may not add to total because

there were 169 children that reported

a race other than White *alone*, Black

alone or White ***and*** Black

Table 2. Race reporting by gender of children under age one in families with a Black *alone* dad and a White *alone* mom

Race reported	Sex of Child			
	Male		Female	
	Number	Percent	Number	Percent
Total	18,038	100.00	17,414	100.00
White *alone*	2,807	15.56	2,901	16.66
Black *alone*	6,495	36.01	5,956	34.20
White ***and*** Black	8,648	47.94	8,476	48.67

Numbers may not add to total because there were 169

children that reported a race other than White *alone*, Black

alone or White ***and*** Black

Table 3. Race reporting by state for children under age one in families with a Black *alone* dad and a White *alone* mom: Sorted alphabetically within region

Region/State	Percent of children reported as: White alone	Black alone	White ***and*** Black
MIDWEST			
Illinoise	13.86	32.96	52.51
Indiana	13.36	32.65	53.42
Iowa	9.01	33.85	57.14
Kansas	11.13	30.42	58.05
Michigan	10.12	29.43	59.45
Minnesota	9.40	29.34	60.78
Missouri	10.62	31.31	58.07
Nebraska	12.06	37.69	49.75
North Dakota	18.75	16.67	64.58
Ohio	11.07	32.16	56.09
South Dakota	12.12	42.42	42.42
Wisconsin	13.77	29.71	56.16
NORTHEAST			
Connecticut	20.97	34.27	44.35
Maine	17.11	30.26	52.63
Massachusetts	27.52	33.98	37.59
New Hampshire	14.58	51.04	33.33
New Jersey	25.65	35.81	37.73
New York	31.00	34.80	33.48
Pennsylvania	14.49	34.72	50.57
Rhode Island	37.97	29.11	32.91
Vermont	16.67	29.17	54.17
SOUTH			
Alabama	18.18	32.99	48.31
Arkansas	16.48	33.33	50.18
Delaware	11.25	32.50	56.25
Washington DC	24.49	28.57	46.94
Florida	25.67	34.71	39.03
Georgia	15.73	34.51	49.68
Kentucky	12.01	32.62	55.02
Louisiana	12.15	48.61	39.04
Maryland	10.56	36.43	52.62
Mississippi	12.00	34.67	53.33
North Carolina	15.70	38.24	45.89
Oklahoma	8.79	35.36	55.02
South Carolina	10.76	34.06	54.78
Tennessee	14.97	31.95	52.67
Texas	14.84	45.64	38.97
Virginia	13.10	34.27	52.26
West Virginia	11.40	37.82	50.78
WEST			
Alaska	10.19	21.30	68.52
Arizona	13.00	36.02	50.67
Californa	16.88	36.44	46.11
Colorado	11.71	33.50	54.63

Hawaii	9.64	32.53	57.83
Idaho	12.96	35.19	51.85
Montana	12.50	21.88	62.50
Nevada	12.18	38.81	48.16
New Mexico	15.85	51.22	32.93
Oregon	10.17	33.22	56.61
Utah	13.25	28.92	57.83
Washington	8.89	25.84	65.14
Wyoming	10.00	16.67	73.33

Numbers may not add to total because there were 169 children that reported a race other than White *alone*, Black *alone* or White ***and*** Black

Table 4. Race reporting of children under age one in families with a Black *alone* dad and a White *alone* mom by percent of White ***and*** Black population in the state

Race of child	State concentration of White ***and*** Black population (percent)			
	0.47 - 0.38	0.37 - 0.29	0.28 - .20	0.19 - 0.11
Total	100.00	100.00	100.00	100.00
White *alone*	24.86	22.11	23.87	22.94
Black *alone*	34.80	35.30	39.02	37.75
White ***and*** Black	39.56	41.76	36.50	38.72

Numbers may not add to total because there were 169 children that reported

a race other than White alone, Black alone or White ***and*** Black

Figure 2. Race reporting of children under age one in families with a Black *alone* dad and a White *alone* mom