PRELIMINARY DRAFT
Do not cite or quote
February 1987

MEASUREMENT OF ETHNIC ORIGIN IN THE DECENNIAL CENSUS*

by

Jeffrey S. Passel and Michael J. Levin

Population Division
U.S. Bureau of the Census
Washington, D.C. 20233
(301) 763-5590

* Paper presented at the 1987 Annual Meeting of the American Association for the Advancement of Science, session on "Whatever Happened to the Melting Pot?" Chicago, Illinois, February 14-18, 1987.

PRELIMINARY DRAFT
Do not cite or quote
February 1987

MEASUREMENT OF ETHNIC ORIGIN IN THE DECENNIAL CENSUS*

by

Jeffrey S. Passel and Michael J. Levin
Population Division
U.S. Bureau of the Census
Washington, D.C. 20233
(301) 763-5590

ABSTRACT

Over the years, the Bureau of the Census has measured the ethnic origin of the population in a variety of ways. The census questions related to ethnic origin with the longest history are those on race (1790) and country of birth (1850). The newest item on ethnicity is the open-ended ancestry question that appeared for the first time in the 1980 census. Other items related to ethnicity from previous and current decennial censuses include race, Hispanic origin or descent, birthplace of parents, citizenship, year of immigration, mother tongue, and language usage. This paper reviews the history of the collection of ethnicity data in decennial censuses. We then focus on race, Hispanic origin, and ancestry data from the 1980 census to discuss the quality of the data and their comparability with the same data items from previous censuses and other data collection systems. Modifications to the race and Hispanic origin data from the 1980 census which correct for reporting errors and differences from historical definitions are described. We draw on data from the November 1979 Current Population Survey supplement on ancestry, the 1980 Content Reinterview Survey, and the 1986 National Content Test to address the accuracy and reliability of responses to the ancestry item in the 1980 census. The paper concludes with a discussion of some of the proposals for collecting ethnicity data in the 1990 census.

* Paper presented at the 1987 Annual Meeting of the American Association for the Advancement of Science, session on "Whatever Happened to the Melting Pot?" Chicago, Illinois, February 14-18, 1987.

MEASUREMENT OF ETHNIC ORIGIN IN THE DECENNIAL CENSUS

Jeffrey S. Passel and Michael J. Levin
U.S. Bureau of the Census

INTRODUCTION

Decennial censuses in the United States provide snapshots of the number, characteristics, and distribution of the population at a point in time. Not only do the results of the various censuses provide a picture of the changing American population, but the content of the census itself provides insight into salient features of American society. The choice of which questions to ask in a decennial census reflects the topics of concern to policymakers and the populace. The way in which questions are asked and data coded shows which categories are judged to be important. As the American population changes over time, the questions and responses change also. The collection of data on ethnic origin in the decennial censuses provides a prime example of this process.

Ethnic origin (or ancestry) is an extremely complicated concept to measure. It is more difficult than most other social or economic characteristics because of the lack of clear-cut definitions, changing terminology, poor reliability, and lack of knowledge on the part of respondents of the degree of affiliation with a group or groups. Some aspects of ethnicity have proved easier to measure because they are more objective and less susceptible to changes in reporting patterns; these would include place of birth, parental birthplace, and mother tongue. Until 1980, the decennial census did not include a general question on ancestry or ethnicity, but rather relied on reported racial identification or the indirect indicators mentioned above to determine ethnic origin.

In the United States, ethnicity or ancestry has a strong geographic orientation because virtually everyone in the population is descended from an immigrant. By ancestry, we generally refer to a person's ethnic group, "roots," or country in which the person, the person's parents, or ancestors were born, regardless of the number of generations removed from the country of origin. Ancestry would reflect identification, but not necessarily the degree of association or attachment with the particular ethnic group(s). As links to the country of origin become more remote in time, the degree of individual identification may lessen and make the measurement of ethnicity even more difficult. Intermarriage serves to further weaken ethnic identification because individuals end up with multiple ethnicities. As a result of factors such as these, an individual's criteria for inclusion in a particular ethnic group are often quite subjective, leading to considerable problems in the consistent measurement of ethnicity.

Decennial censuses in the United States have used four types of questions to describe the cultural and geographic origins of the population. First, data on race have been collected in every decennial census dating back to the first one in 1790. (Although the form of the question and the response categories have changed considerably.) Second, a question about birthplace has been included in the decennial censuses since 1850, allowing the determination of immigrants' countries of origin. Censuses from 1870 through 1970 also asked respondents about their parents' birthplaces, so that second-generation and third-or-higher-generation Americans could be identified. Third, censuses have included questions about language usage, such as mother tongue, ability to read English, and the use of languages other than English. Questions of this type were first asked in 1890, but

tabulations of data on foreign mother tongue date from 1910. Finally, the fourth type of question is a direct inquiry about self-identified ethnicity or ancestry. The 1970 and 1980 censuses both included a question about Hispanic (or Spanish) origin, but the 1980 census was the first to include a general, open-ended question asking respondents to report ancestry directly. Table 1 summarizes the questions related to ancestry or ethnicity in censuses since 1850.

This paper explores some of the problems of measuring ethnic origin in the decennial census. Specifically, we focus on three questions used in the 1980 census to measure ethnic origin: the race question, the Hispanic/ Spanish origin question, and the ancestry question. For all three of these, we analyze data from the 1980 census and discuss plans for 1990. We treat the race and Hispanic origin questions principally in terms of reliability and historical comparability. For these two questions, the main issues relate to anomalies in the 1980 census data and how to deal with them. (See also Passel and Word, 1987 for detailed treatment of "corrections".)

The data on self-reported ancestry are treated in more detail as part of the process leading to the development of questions for the 1990 decennial census. Since the items on parental birthplace and mother tongue were eliminated from the 1980 census, analysis of the quality of self-reported ancestry has been made more difficult. In this paper, we draw on data from a November 1979 supplement on ancestry to the Current Population Survey (CPS), the 1980 census itself, the Content Reinterview Survey (CRS) for the 1980 census, and the 1986 National Content Test (NCT) to address issues relating to the quality of data on ancestry.

RACE

1980 Census Data as Collected

Data on racial groups as collected in the 1980 census and published in census publications have proved difficult to use for many persons who try to combine the data with historical census data or data from other collection systems. The 1980 census results had a large number of persons (about 6.8 million) in a residual race category which was designated "Other, not specified." (See Figure 1 for a facsimile of the 1980 question and Table 2 for the data.) The vast majority of these persons were of Hispanic origin and most would have been classified as "White" in previous censuses and in many other data systems. Another large group in the residual category consisted of Asians and Pacific Islanders whose specific nationalities were not listed on the census form.

Inconsistencies between the 1980 census and previous censuses in racial definitions distort historical comparisons, particularly for Whites and for Other races. Lack of comparability between 1980 census data and vital statistics can lead to serious distortions in vital rates, especially for these same two racial groups (Chilton and Sutton, 1986). The distortions are noticeable at the national level, but can be extremely serious for local areas, particularly those with sizeable Hispanic populations.

Preliminary OMB-Consistent Modified Race Data (1980 Census)

The Census Bureau has recognized the inconsistencies between racial classifications used in the 1980 census and other definitions. For its own postcensal population estimates and for outside data users, the Census Bureau produced "Preliminary OMB-Consistent Modified Race Data" (Passel, 1982). The

alternative race data were based on cross-tabulations by age, sex, race, and Hispanic origin from the 1980 census complete count at the state and county level. The construction of the preliminary modified race data also used information on write-in responses to the race item in 1980 census sample data at the state level. The modified race data had 6.34 million more Whites than the initial census figures, or 3.4 percent; 188,000 more Blacks, or 0.7 percent; and 6.53 million fewer persons of Other race, or a reduction of 55.9 percent. (See Tables 2 and 3.)

The "Preliminary OMB-Consistent Modified Race Data" suffer from three major limitations. First, the only tabulations available are for age, sex, race (White, Black, and Other races only), and Hispanic origin. Because they are based on aggregated data, it is difficult (or impossible) to produce consistent tabulations on characteristics other than age, sex, race, and Hispanic origin. Geographic detail is limited to the nation, states, and counties. Finally, for some of the modifications, state-specific factors were used to produce county-level data. As a result, the modifications may not be highly accurate in some counties.

Micro-Modified Race Data (1980 Census)

To remedy the shortcomings of the "Preliminary OMB-Consistent Modified Race Data," the Census Bureau has developed what are called "Micro-Modified Race Data" from the sample data in the 1980 census. All of the 5.95 million persons in the sample who reported in the "Other, not specified" race category were assigned to specific racial groups on the basis of other data from the census related to ethnicity, by household relationships, and eventually by a "hot deck" procedure. Virtually all of these individuals were assigned to one of the four OMB-defined race groups: White, Black,

American Indian or Alaska Native, and Asian or Pacific Islander.[1/] A very small number of persons (about 1,000) remained in the "other" category. (The explanation which follows is a brief summary of the work described in detail in Passel and Word, 1987.)

To make the reassignments, each person of "Other, not specified" race was subjected to a series of hierarchical edits using other census data related to racial identification:

1. Direct Assignment by **RACE**
2. Direct Assignment by **ANCESTRY**
3. Direct Assignment by **MEXICAN ORIGIN**
4. Assignment by Household **RELATIONSHIP**
5. "**HOT DECK**" Assignment

For direct assignment by race, certain specific write-in responses to the race question were assigned to one of the four OMB categories, for example: Laotian to Asian or Pacific Islander, Argentinian to White. These edits assigned 3.39 million persons or 56.9 percent of the original responses of "Other, not specified" race to specific racial groups. (See table 5.) Almost all of these persons (3.18 million) were assigned as White; about 2.95 million of these persons had responded "Mexican" to the race question.

After the direct assignment by race, about 2.56 million persons remained in the "Other, not specified" category. Most of the remaining persons (70 percent) had given a Hispanic response to the race question (such as, "Puerto Rican" or "Hispano") that could not be assigned unequivocally to a racial group. Another 28 percent had responded as "Other," but had not provided a written response. The next set of edit rules made assignments to racial groups on the basis of responses to the ancestry question. For example, persons of Italian ancestry were assigned as White; persons of Afro-American ancestry, as Black; persons of Japanese ancestry, as Asian.

The ancestry set of edit rules assigned 948,000 persons, or 36.9 percent of those remaining. Most of these persons, 843,000 or 88.9 percent were assigned as White.

The third set of edit rules assigned 132,000 persons who were not yet classified but were of Mexican origin to the White category. The fourth set of edits was based on household relationships. Individuals who still remained in the "Other, not specified" category, but were related to persons with a specified race were assigned to the appropriate OMB racial groups. The rules based on household relationship assigned 325,000 persons; of these, 265,000 or 81.7 percent were assigned as White.

Even after applying the myriad of edit rules to the original 5.95 million persons in the "Other, not specified" category, a sizeable number of persons (about 1.16 million) could still not be classified. For example, many persons of Puerto Rican origin who responded as Puerto Rican to the race question could not be assigned as either White or Black by the edit rules. Another large group which remained unclassified at this point were persons who marked the "Other, not specified" category, but did not provide a written response. The remaining unclassified persons were assigned to racial groups with a "hot-decking" procedure stratified by type of Hispanic origin.

Overall, the micro-modification procedures added 5.45 million persons, or 2.9 percent to the White population in the 1980 census. (See Table 4.) The modified Black population was increased by 216,000 or 0.8 percent; the American Indian and Alaska Native population increased by 16,000 or 1.0 percent; and the Asian and Pacific Islander population by 265,000 or 7.5 percent. The principal difference, at the national level, between the

micro-modified race data and the preliminary OMB-consistent data is that the Other races category in the micro-modified data is larger by 209,000 or 3.9 percent. A major portion of this difference (about 156,000) can be traced to differences between the unmodified sample and complete counts for Asians and American Indians. (See Table 2. Passel and Berman, 1986 discuss these differences for American Indians in detail.)

One major strength of the micro-modification procedure is that there is now a file of all persons in the 1980 census sample that includes a racial designator which is consistent with the OMB definition and most other data systems. From this file, it is possible to produce tabulations for any variable and for any geographic area. Also, because the preliminary OMB-consistent procedure used state-specific factors at the county-level, corrections in the micro-modified data should be more accurate for most substate areas. However, because the file contains sample data, figures for small areas or small categories may be subject to large sampling errors. Passel and Word (1987) include a detailed description of the edit rules, tables showing the workings of the edit process, and modified racial data for subnational areas.

Prospects for 1990

For 1990, two versions of the race question are being considered. (See Figure 1.) Both versions are designed to collect data on all Asians and Pacific Islanders in the complete count, thus eliminating the initial shortfall in this category in the 100-percent data. The word "Race" has been reinstated as part of the question in both versions. These two forms of the question are being tested to see if they induce more persons of Hispanic origin to respond with a specific race. (See McKenney, 1986 for a discussion of the 1990 testing program for racial and ethnic items.)

In any case, it is virtually certain that a large portion of the Hispanic population will continue to respond in the Other race category in 1990. As a consequence, the race data from the 1990 census will again be inconsistent with data from censuses before 1980 and with data from many other systems.2/ Thus, to produce consistent and meaningful results by race when comparing 1990 census figures with historical data series and other data systems, it will probably be necessary to modify the 1990 census data in a manner similar to what has been done for 1980. This situation is almost certainly unavoidable when data are collected by self-enumeration and self-identification in a society such as the United States' where racial and ethnic identities are not fixed and where the concepts of race and ethnicity mean different things to different people.

HISPANIC ORIGIN

1970 Census Data and Other Early Attempts

The Census Bureau first collected data on the Spanish origin population on a national scale in the November 1969 Current Population Survey. This was followed by the 1970 census which counted 9.1 million persons of Spanish origin on the basis of a self-identification question included in the 5-percent sample. (See Figure 2.) Various evaluations of the 1970 data showed the Spanish origin results to be rather poor. Only 75 percent of persons reporting Spanish origin in the 1970 CRS had answered affirmatively to the census question. In addition, various other response problems occurred, including massive misreporting into the Spanish origin category, as well as apparently serious nonresponse and erroneous omissions from the Spanish origin population. (The material in this section is drawn principally from Siegel and Passel, 1979.)

The "Central and South American origin" category in the 1970 census had a population of 1.51 million persons. By way of comparison, the November 1969 CPS had 556,000 persons of Central and South American origin; the March 1971 CPS had 501,000; and the March 1972 CPS, 599,000. (See Table 6.) These data suggest that the 1970 census had almost 1 million too many persons in this category. The geographic distribution of the Central and South American origin population showed large concentrations in the central and southern United States. Many people apparently interpreted "Central and South American" origin to mean being from the central or southern United States. Taken together with the results from the 1970 CRS, the 1970 census and the CPS results point to an extremely large number of erroneous responses on the part of non-Hispanic persons.3/

Other analyses of 1970 data suggest, however, that the Spanish origin data seriously understate the number of Hispanics in the census. Several different "Hispanic" populations were constructed from 1970 census data with other questions that could be used to ascertain ethnicity. These populations include: Spanish language, Spanish heritage, Spanish language or surname, Spanish birth or parentage, and Spanish surname. The largest of these was the Spanish language or surname population at 10.1 million (Table 7). In parts of the country with concentrations of persons of Hispanic origin, the Spanish language population and the Spanish language or surname population showed substantially larger numbers than the Spanish origin population.

A further example of the shortfall of the 1970 census is shown in Table 6. The Mexican origin population in the census was 4.5 million. In the November 1969 CPS, the March 1971 CPS, and the March 1972 CPS, the Mexican origin population exceeded 5 million, or about 10 percent more than the 1970 census.

In addition, the CPS estimates of the Cuban and Other Spanish populations were also significantly higher than the 1970 census. Thus, even though the Spanish origin population in the 1970 census included almost 1 million too many persons of Central or South American origin, it appears that the population was severely understated because many Hispanic persons who were counted in the 1970 census did not respond affirmatively to the Spanish origin question.

Data collected using the Spanish origin identifier during the 1970s did not appear to suffer from the same deficiencies as the 1970 census data. Reinterview studies conducted in conjunction with test censuses, year-to-year match studies of the Current Population Survey, and the 1976 NCT all showed that the Spanish origin question produced consistent data. Roughly 90 percent or more of persons who responded as Hispanic in reinterviews had responded as Hispanic in the original interview.

The improvement and general acceptance of the Spanish origin question during the 1970s set the stage for its inclusion as a 100-percent item in the 1980 census. Furthermore, the continued improvement in data quality suggests that the 1970 census figures on the Spanish origin population should **not** be used in historical analyses. A better approximation from the 1970 census to Hispanic data collected after 1970 is the Spanish language identifier.

1980 Census Data as Collected

Data from the 1980 census for the population of Hispanic origin have been shown to be of high quality. The 1980 CRS showed that approximately 90 percent of persons reporting Hispanic origin in the census did so in the reinterview, and vice versa (Fernandez, 1986). This level of consistency,

which is consistent with historical experience after the 1970 census, basically reflects the subjective nature of the identifier (Siegel and Passel, 1979). As in the past, the "Other Spanish" category proved to be the most inconsistent group. (See Figure 3 for a facsimile of the question.) Tienda and Ortiz (1986) analyzed Hispanic origin data from the 1980 census in conjunction with other items related to ethnicity. They also found substantial agreement among the items and concluded that the Hispanic origin question proved to be the best single item to identify the Hispanic population.

In spite of the overall high quality of the 1980 data on the Hispanic population, the 1980 census did show an unexpectedly large number of persons of Mexican origin in areas which historically have not had large Mexican populations, notably eastern and southern states. Furthermore, an unexpectedly large proportion of the Mexican origin population in many areas was Black. Early tabulations of sample data showed that a large proportion of these persons did not possess characteristics normally associated with Mexican ethnicity. A follow-up survey conducted to investigate this anomaly found that virtually all of the persons suspected of misreporting (over 200,000) did not consider themselves either Mexican or Hispanic. As with the "Central or South American origin" category in the 1970 census, respondents apparently misunderstood the intent of the Hispanic origin question and believed that they were responding as "American" when they marked the category "Mex.-Amer." (See U.S. Bureau of the Census, 1982a for the details of the study.)

Further analysis of the Hispanic data suggested that similar misreporting occurred in the same areas for the "Other Spanish origin" category. On the

other hand, tabulations of sample data from the 1980 census showed that about 85,000 people who did not report as Hispanic origin had three or more of the following characteristics which would normally be associated with being of Hispanic origin: Hispanic racial response, Spanish spoken in the home, Hispanic ancestry, born in a Hispanic country, and Spanish surname (5 southwestern states only). Although none of the analyses showed a massive amount of misreporting (either into or out of the Hispanic origin population) at the national level, they did indicate the possibility of substantial distortions at the local level.

Micro-Modified Hispanic Origin Data (1980 Census)

To measure the degree of potential misreporting and to correct the data, a procedure for evaluating and modifying the responses to the Hispanic origin item in the 1980 census was developed which analyzed responses to census questions on place of birth, language, race, ancestry, language spoken in the home, and residence 5 years ago. Passel and Word (1987) examined individual records from the sample detail file for all persons who responded as Hispanic origin plus any person with a Hispanic response to any of the items which could indicate ethnicity (race, ancestry, country of birth, language, or surname) as well as all other persons in their households. Every person in these households was assigned a score on a "Hispanicity" index which was based on responses to the ethnicity items, the current state of residence, and the state or country of birth. The more of the characteristics identifiable as Hispanic that are possessed by the individual, the higher the Hispanicity index. The specifics of the scoring system are shown in Table 8. (See Passel and Word, 1987 for more details and further explanation of the edit rules, which are summarized in the remainder of this section. See also Word, 1984.)

The quality of each individual's response to the Hispanic origin item was evaluated on the basis of the Hispanicity score taken alone, plus taken in the context of the scores and responses of other household members. Each individual was assigned a "modified Hispanic origin" on the basis of the original response, the ethnicity score, and the scores and responses of other household members. The basic strategy in this modification was to accept the original response unless there were very strong indications that the response was erroneous. Virtually all original responses (both Hispanic and non-Hispanic) were retained as the modified Hispanic origin category.

Individuals were retained as Hispanic origin if <u>any</u> response to <u>any</u> of the ethnicity items suggested that the person could be Hispanic.[4/] In addition, if any member of the household was retained as Hispanic, then all members of the household who responded originally as Hispanic were also retained as Hispanic, regardless of their ethnicity scores. Thus, the only way a person could be changed from Hispanic origin to non-Hispanic origin was if <u>no</u> household member had <u>any</u> Hispanic ethnic characteristic. At the national level, only 581,000 persons were changed from Hispanic origin to non-Hispanic origin. Overall, 96.0 percent of the Hispanic origin population was retained as Hispanic. (See Table 9.) However, for many areas within the country, the proportion retained was considerably smaller; for example, in the East South Central division, only 45.0 percent of the original Hispanic origin population was retained as Hispanic.

Persons who had not responded as Hispanic could be added to the micro-modified Hispanic population if their ethnicity score provided very strong evidence that the original non-Hispanic response was erroneous. For example, any two of the following would provide such evidence: a write-in

Hispanic race entry, Hispanic ancestry, a Hispanic country of birth, Spanish language spoken in the home (except the combination of place of birth and language). Other qualified combinations, i.e., those leading to an index score of 1.8 or more, can be derived from Table 8. For all additions to the modified Hispanic population, persons added were assigned to specific Hispanic categories.

In adding persons to the micro-modified Hispanic population, household relationships also played a major role. In households where some persons had originally responded as Hispanic (i.e., a mixed Hispanic/non-Hispanic household), very strong evidence was required to add a non-Hispanic person to the Hispanic origin population on the theory that the household members understood the intent of the question. As previously described, a score of 1.8 or more for a persons who had originally responded as non-Hispanic in a mixed household was required to effect a change into the modified Hispanic population.

In households where no one had originally responded as Hispanic, the first addition to the modified Hispanic population required the very same strong evidence as in mixed Hispanic households (i.e., a score of 1.8 or more). Other household members after the first one added could be added to the modified Hispanic population on lesser evidence (i.e., a score of 1.0 or more).

The edit procedures for adding to the Hispanic population shifted about 228,000 persons into the modified Hispanic population. This figure represents about 1.6 percent of the modified Hispanic population (Table 9). In fact, for the 9 states with the largest Hispanic populations--states which

account for over 85 percent of the Hispanic population, the proportion of persons added to the modified Hispanic population was exactly 1.6 percent. The rate was the same for the remaining 41 states and the District of Columbia.

The overall net effect of the modifications was to reduce the sample figure for the Hispanic population to 14.3 million or by 353,000, that is, 2.4 percent of the original sample population. Because the misresponse into the Hispanic origin tended to be in areas without large Hispanic populations, the modification increased the geographic concentration of the Hispanic population. Thus, the 9 states with the largest Hispanic populations have 87.1 percent of the modified Hispanic population as compared to 85.1 percent of the original sample Hispanic population in the 1980 census.

Results of the modification of each type of Hispanic origin (Table 10) are consistent with the evaluations from the 1980 CRS and earlier. The Puerto Rican and Cuban origin populations are highly consistent; the Mexican origin population, less so; and the Other Spanish population, even less. The modifications had very little effect on the Puerto Rican and Cuban origin populations as the modified population differed by less than one percent from the original figures. The rate at which persons were added to the population was roughly the same in all parts of the country. Deletions were, however, higher in parts of the country with few Puerto Ricans or Cubans.

Overall, the Mexican origin data also proved to be of high quality as the rate of addition was only 1.0 percent and of deletion, 3.1 percent. The geographic distribution of the deletions reflects the results of the earlier study (U.S. Bureau of the Census, 1982a) which found substantial misreporting

in eastern and southern states. The addition rate was again roughly the same everywhere. In the 6 states with the largest Mexican origin populations, there were moderate to small increases in the modified population as the deletions approximated the additions. In the rest of the country, about 1 in 5 of the original responses were assessed to be in error. (In some areas, such as the East South Central division, more than half of the original responses were modified.) The differential rate of deletions led to a greater geographic concentration of Hispanics in the micro-modified population. Whereas 87.5 percent of the original Mexican origin population lived in the 6 states with the largest numbers, 89.7 percent of the modified population did so.

The Other Spanish origin population has previously proved to be less consistently reported than other Hispanic origin groups. The micro-modified data show this same pattern. The addition rate for this modified group was about 3 percent in all areas--a rate which exceeds the other origin groups.[5/] The rate of deletion in the states with the largest numbers of Other Spanish approached 5 percent and was greater than the corresponding rate for other groups. In states with smaller numbers of Other Spanish, the deletion rate approached the 1 in 5 value found among the Mexican origin population. Although no direct studies of misreporting were done for the Other Spanish population as they were for the Mexican origin group, the high deletion rate suggests the presence of response problems like those that occurred in the original Mexican origin responses.

This modified Hispanic origin indicator is now appended to the record of every person in the sample detail file of the 1980 census.[6/] From the modified sample detail file, tabulations of modified Hispanic origin, as well

as modified race, for any variable of interest can be produced for any geographic area, subject to the limitations described previously for the micro-modified race variable. Passel and Word (1987) include a description of the edit rules and scoring system used in constructing the modified Hispanic origin identifier, tables showing the results of the modification process, and modified Hispanic data for subnational areas. Although a great deal of attention was directed to the modification process for the Hispanic data, it should be stressed that the unmodified data are of sufficient quality for most analyses, especially in areas with large Hispanic populations.

Prospects for 1990

Two slightly modified versions of the Hispanic origin question are being tested for use in the 1990 census (Figure 3). Both versions would provide more detailed information on the specific origins of the "Other Spanish" population in sample data. The short format also might help eliminate misreporting into the "Mexican-American" category because it is not listed as a precoded box. However, this version would make it more difficult to obtain 100-percent data on the Hispanic population by type. Both versions are being tested to see if they can continue to produce the high quality data on the Hispanic population available from the 1980 census (McKenney, 1986)

ANCESTRY

When the place of birth question was first asked in 1850, about 10 percent of the population was foreign-born. A high volume of immigration followed the Civil War, causing the proportion foreign-born to increase, peaking in the early decades of the 20th century (Table 11). Restrictive immigration laws

of the 1920s led to decreases in both the number and the proportion foreign-born after 1930. With the revision of immigration laws in the 1960s, immigration increased considerably and the origins of the immigrants shifted. Latin America and Asia supply a far higher proportion of immigrants today than 25 years ago.

The proportion of second-generation Americans, that is, native-born persons with 1 or 2 foreign-born parents, lags several decades behind the portion foreign-born. By 1920, 21 percent of the population were second-generation, but the proportion has decreased steadily since then. If the foreign-born population continues to grow more rapidly than the native, the long-term decline in second-generation Americans may be reversed (depending of course on the fertility rates of the native and foreign-born populations).

Data on foreign-born and foreign-parentage (i.e., the first and second generations) only describe immigration trends of a few decades. Any effect of immigration in the more distant past on ethnic composition is not measured by the birthplace and parental birthplace items when the immigrant ancestors are grandparents or more distant. The 1980 census differed from all other censuses since 1870 because it did not explicitly provide information about the sons and daughters of the foreign-born, but it did provide more general ancestry information. Since the ancestry question is an innovation and may be the most appropriate question to identify a large segment of the population, we will considered in detail the usefulness and reliability of the data.

Ancestry Data as Collected (1980 Census and Recent Surveys)

The ancestry question in the 1980 census was based on self-identification, was open-ended, and had no prelisted categories. (See Figure 4.) The November 1979 CPS and the 1986 NCT also collected ancestry data with essentially the same question.[7] Some individuals reported a single ancestry group, others reported more than one group. All single- and double-ancestry responses were coded in each survey and the census.[8] In addition, 17 triple-origin ancestries expected to be frequently reported were coded, while only the first two reported ancestries were coded for all other responses of three or more ancestries.

Differences between the census and the two surveys in the method of data collection can have great impact on the data. Self-enumeration was used in both the census and the NCT. Even though ancestry was based on self-identification in the CPS also, selection of response categories was not completely independent since enumerators were instructed to prod respondents for a specific ancestry if a religion, the category "American," or an unclassifiable response was given. The census and the NCT each provided a number of examples (different examples in each case), such as Hungarian, Irish, Italian, English, and Afro-American, to aid respondents; the listing may have influenced the reporting as well. In addition, the instructions on the NCT requested respondents to write "the group" with which they identified and probably led to increased reporting of single ancestries at the expense of multiple ancestry reporting.

Approximately the same proportions of the population reported as "American" or "United States" in each survey and the census, even though interviewers were instructed in the CPS to explain that "American" or a religious response

was not appropriate, since ancestry was to refer to the specific foreign nationality of the person or his or her ancestors. On the other hand, the percentage reporting a religion or other unclassifiable response was much less when enumerators were used to collect the data in the CPS. The proportion of persons not reporting any ancestry response remained fairly consistent in the three data collection operations. The proportion of nonresponse was slightly lower in the CPS than in the census. The proportion apparently increased to 13 percent in the NCT, but this difference may not be statistically significant because of the small number of cases in the NCT.

German was the largest ancestry group in the CPS and the NCT, but English was most frequently selected in the census (Table 13). The 1980 census questionnaire design may have contributed to the differences in ancestry reporting. The prominence of the term "English" in the census question on language (which immediately preceded the ancestry question) and the listing of "English" as the second example in the ancestry question may have influenced respondents to report a single ancestry of English. The contribution of single English ancestry is also seen in the large proportion of single ancestry in the census seen in Table 12. In the NCT, when the language question did not appear with the ancestry question, and when English was not on the list of examples, the percentage reporting single and multiple ancestry for English ancestry was similar to the CPS. The contribution of single English ancestry is also seen in the large proportion of single ancestry reporting in the census seen in Table 12. (Note that because persons who reported multiple ancestries were included in more than one group, the sum of persons reporting the ancestry groups was greater than the total; for example, a person reporting "German-English" was tabulated in both the "German and other group(s)" and "English and other group(s)" categories.)

What Does the Ancestry Question Measure?

The ancestry question itself, a direct measure of ethnicity, would seem to be a substantial improvement over previous indirect measures (place of birth of person and parents and language) for identifying ethnic groups. First, a major advantage of the ancestry question is that it allows identification of the ethnicity of all persons, not just persons who would be classified on the basis of their own birthplace or birthplaces of their parents. Second, unlike birthplace, ethnicity provides data on the individual's own perception of ethnic identification. Third, the ancestry question is much more informative than the language inquiries.

The ancestry question differs from many other socio-demographic queries since responses may not be "correct" in the same manner as responses to questions about age or income. Although the accuracy of a reported ancestry cannot be determined exactly, some of the factors which are related to ancestry can be studied. The CPS, the census, and the NCT each gathered information to assist in this analysis. The 1979 CPS collected information on five items that could which influence a person's selection of an ancestry--birthplace, mother's birthplace, father's birthplace, mother tongue, and current language. The census gathered data on race, Hispanic origin, birthplace, and language (which can be used primarily in assessing data for Asians and persons of Hispanic origin). The NCT gathered information only on own and parental birthplace[9/], but the NCT provided additional information because the reinterview asked the ancestry questions again and the responses were matched in processing.

Several attempts have been made to develop correspondences between direct and indirect measures of ancestry by looking at generation since immigration.

For example, Levin and Farley (1982) developed correspondences between ancestry and countries of birth and languages for ancestries in their analysis of the November 1979 CPS. They found that in some situations, correspondences were easily obtained. For Korean ancestry, for example, Korea was the appropriate country of birth, and Korean was the language. Some ancestries were associated with two or more languages, such as both French and Dutch for persons reporting Belgian ancestry. Other ancestries such as West Indian or Slavic involve many countries of birth. In addition, it can be difficult to specify countries and languages for many Central and Eastern European ancestries because there have been numerous border changes and the same language is spoken in several countries.

The correspondences were used to determine generation of immigration. For example, an American-born person who reported as Italian and whose father was born in Italy became a second-generation Italian. Those persons whose birthplace or parents' birthplace did not agree with the reported ancestry were classified as third or later generations.

Table 15 presents selected ancestry groups (single and multiple combined) with more than 100,000 persons in the November 1979 CPS and illustrates one of the advantages of the ancestry question. The next-to-last column shows the proportion of the population that was third-and-later generation; in fact, the table is arranged in descending order of this percentage. Information about a person's birthplace and parental birthplaces would only identify a small portion of some major ethnic groups. Less than 7 percent of the people who actually reported African, Irish, Dutch, Scottish, Welsh, English, Scandinavian, German, and French ancestries would be classified in those groups in 1979 if only birthplace questions were asked. Thus, the

ancestry question provides a more complete measurement for identifying these groups, especially for those whose forebears came to the United States in the 19th or early 20th centuries. On the other hand, this question is no more advantageous than the birthplace questions for identifying some of the groups that are recent arrivals, such as Korean, Iranian, Cuban, Vietnamese, or Colombian. In these instances, ancestry measures little more than the first and second generations.

It is important to point out that, although the ancestry question clearly has the potential for providing ethnic data for a very large proportion of the population, it is not clear how reliable these data are. Furthermore, persons reporting multiple ancestries may not be adequately considered in this analytic scheme. After all, if a person reports as German-English ancestry and has one parent born in Germany, the person has agreement at the second generation with one ancestry, but may or may not have agreement with the other.

Reporting of multiple ancestries can also make analysis and understanding of the data problematic. The proportion of third and subsequent generations who claim multiple ancestries illustrates the relative levels of intermarriage across ethnic groups and the consequent difficulties in measurement. (See the last column in Table 15.) In order to claim multiple ancestry, in theory, a person's parents or earlier ancestors must have been of different ancestries. In other words, in order to have a multiple ancestry, a marriage of two single ancestries should have occurred at some point. For some groups--Asian Indian, for example--about half of the persons born in the U.S. and whose parents were both born in the U.S. claimed multiple ancestries, that is, Asian Indian and some other ancestry. On the other hand, about