

**Accuracy of Data Collected
by the Census Question on
Immigrants' Year of Arrival**

by

Dowell Myers

January 2004

Working Paper PDRG04-01

Population Dynamics Research Group
School of Policy, Planning, and Development
University of Southern California

www.usc.edu/schools/sppd/research/popdynamics/

The great majority of our knowledge about foreign born residents or immigrants¹ in the United States comes from the decennial census or the Current Population Survey (CPS). Both data collection programs ask respondents' place of birth and, for foreign born, year of arrival in the U.S. (Costanzo et al. 2002). The decennial census has proven especially valuable for research on immigrants because its public use microdata files provide a 1-in-20 sample of the population. In contrast, the CPS sample is fewer than 1-in-1000. The large size of the sample from the decennial census is useful for analysis of small groups defined by country of origin or by detailed age and other statuses.

This brief offers a summary evaluation of the accuracy of data collected by the census question on year of arrival. This question also provides the basis for measures of immigrant duration, or length of time residing in the United States. Presented first is a brief description of the data and its history. Next we address the reasons for concern that have been expressed about the quality of data collected by the census question. Most of the concerns that have been raised are circumstantial but have legitimate basis. Available evidence is then reviewed on the quality of the data, the stability of measurements over time, and the validity of analytical measurements based on the year of arrival data. Conclusions reached are that the year of arrival data, despite some inconsistencies, are technically more reliable and less error prone than other data commonly relied upon. The question on year of arrival also provides a valid instrument for defining arrival cohorts because measures of these cohorts prove remarkably stable over intervals of 20 years.

HISTORY OF THE YEAR OF ARRIVAL QUESTION

Questions on the year in which foreign born residents came to the U.S. to live were asked in the decennial censuses of 1900 to 1930 (Gibson and Lennon 1999; U.S. Census Bureau 2002). Early in the century the question asked: "Year of immigration to the United States." As described below, the question wording was changed in 1970 and again in 2000.

After the 1930 census, the question was not asked again until 1970 (through 2000). Apparently, the lapse in the question's inclusion followed the steep decline in immigration following the acts of 1924 and 1925. The revival of the question was planned in the late 1960s following the Immigration Act of 1965. The exact wording of the question in 1970 was: "For persons born in a foreign country.....When did he come to the United States to stay?" This was followed by nine

¹ The terms immigrant and foreign born resident are used interchangeably in this report. Even though "immigrant" is a legal status, as well as an expression of intention of permanent residence, this distinction cannot be observed with census data, and so it is commonly ignored in studies of the foreign born population.

check-off categories, ranging from “1965 to 70” to “before 1915.” The year of arrival question was preceded by a write-in question asking for place of birth and a citizenship and naturalization question. Identical questions about year of arrival, also preceded by place of birth and citizenship questions, were asked in 1980 and 1990, but response categories were updated to reflect more recent intervals.

In Census 2000, the question was substantially revised. Following place of birth and citizenship questions, the year of arrival query read: “When did this person come to live in the United States?” One change in the question is that it asks “come to live in the United States” instead of “come to the United States to stay.” A second difference from 1970, 1980, and 1990 is that the question used in 2000 asks for a write-in of exact year of arrival instead of a check-off from a list of categories. Yet a third difference is that the year of arrival question was followed immediately by the long-standing migration question on place of residence five years ago. In 1990, there was an interruption between the year of arrival and migration questions. In between these two, respondents were asked about current school attendance, completed educational attainment, and ancestry. In 2000, the two questions on immigration and migration are abutting. Moreover, the migration question has inserted in its opening panel a check-off category, “No, outside the United States,” which is followed by a request to write-in the foreign country of residence. This revision in question sequencing may help to reduce inconsistent responses to the immigration and migration questions (Ellis and Wright 1998) as discussed below.

In broad terms, the content of the census questionnaire is shaped by the perceived importance of different factors. The questions related to immigration respond to current political events and the resulting prevalence of behavior. Because immigration was curtailed in the 1920s, the question on year of arrival was dropped after the 1930 census, and when immigration barriers were formally lowered in the 1960s the question was reintroduced in the 1970 census. Similarly, a question on the nativity of parents was intended to measure the status of the second generation (children of immigrants). That question remained on the census questionnaire up to 1970, after which time the prevalence of second generation sank so low that it was removed. The Current Population Survey now asks questions pertaining to both the immigrants and the second generation.

REASONS FOR CONCERN ABOUT THE QUALITY OF THE DATA

The quality of the data collected by the year of arrival question has come under some criticism. There appear to be two main bases of concerns about the accuracy and meaningfulness of the year of arrival data. The first is that the data have become embroiled in disputes over political and ideological interpretations because the data have been applied to contentious questions of immigration and assimilation. The second is a growing body of evidence on the complexities of migration behavior—particularly emigration and circularity—that obviate the meaning of “come to stay” in the U.S.

Each of these sources of criticisms deserves to be addressed. Following that we will inquire into the relative validity of the respective critiques.

Political Implications of the Data

The question on year of arrival has led to some of the most contentious research about immigrants because data collected by this question have been used to interpret the prospects for assimilation of immigrants. This application has great political implications because conclusions of easier assimilation would support lowering barriers to immigration while the opposite conclusion lends support to greater restrictions. In fact, all that is measured by the year of arrival question is the period when respondents entered the U.S. and the length of time they have resided in the U.S. It is the interpretations of those data that create the controversy.

The debate between Chiswick (1978) and (Borjas (1985) was the first to publicize this intellectual dispute. The two scholars interpreted the same year of arrival information with regard to the “quality” of immigrants, reaching opposite conclusions, because one analysis was cross-sectional and the other cohort longitudinal. The 1970 census was the first in some decades to ask immigrants when they had entered the U.S. With those data, Barry Chiswick (1978) was able to investigate the cross-sectional relationship between Years Since Migration and earnings of immigrants. Chiswick described *differences* in earnings across categories of arrival dates as income *growth*. After the 1980 census provided a second cross-section with earnings and arrival data, George Borjas (1985) was able to observe cohorts of immigrant arrivals at two points in time. He tested whether their earnings growth over the 10-year period was the same as implied by the cross-sectional difference observed in 1970 alone.

Chiswick found that immigrants began their U.S. careers with earnings well below the native-born but after 15 years in the U.S. their earnings had surpassed those of the native-born. He attributed immigrants’ apparent overtaking of the native-born to their superior qualities (more innate ability and stronger motivation). In contrast, Borjas showed that the longer-settled immigrants in 1970 and 1980 had been tracking on higher earnings trajectories all along, while more recent arrivals were tracking on lower trajectories. Thus, a cross-sectional relationship linking earnings of recent arrivals and longer settled immigrants exaggerated the amount of earnings growth to be expected. The cross-sectional relationship summed the effects of actual earnings growth over time for cohorts plus the positive gap between trajectories of high-tracking early arrivals and low-tracking recent arrivals. In contrast, Borjas attributed recent immigrants’ lower tracking to the opposite cause, namely their lower “quality” produced by less selective immigration policies in recent years and by a shift from European to Latin American sources of immigration.

The political implications of the two contrasting analyses could not be starker. One found that immigrants were high quality and would even surpass the native born after admission into the United States. The other found that immigrants were low quality and were stratified into permanently lower status. Advocates for and against immigration seized on

the results, while many scholars remained skeptical of the interpretations drawn on both sides. In this skeptical climate, doubts were cast upon the underlying census data on year of arrival.

In all of the above, it is the application and interpretation of the census year of arrival data that has been questioned. In the attacks on respective interpretations, the data themselves have been cast under a pall of controversy. Indeed, there is a clear basis for questioning how well the year of arrival data capture realities of immigration.

Complexities of Migration Behavior

A separate group of scholars has raised questions about the quality of year of arrival data because of their research on migration behavior. The very meaning of “come to stay” has been challenged. The ideal respondent to the year of arrival question made a single, permanent move to the U.S. This is best typified by the experience of immigrants from Cuba and least typified by those from Mexico. Emigration and circular migration are two aspects of migration behavior that introduce potential bias in the year of entry variable.

The Issue of Selective Emigration

Emigration, or the outmigration of previous immigrants, is estimated to amount to some 195,000 persons per year, or roughly 20% of the volume of immigrants (Ahmed and Robinson 1994). Emigration has the potential to change the quality of immigrant cohorts surviving over time, so that what looks like adaptation might be merely an artifact of “unsuccessful” migrants returning home. Borjas and Bratsberg (1996), comparing 1980 Census data with data from the Immigration and Naturalization Service (INS), found that emigration was highly selective. Migrants were more likely to return to their home country if their home country was wealthy and close to the U.S. A doubling of per-capita GNP increases the out-migration rate by 4.9 percentage points, and every 1,000 mile increase in distance between the U.S. and the sending country decreases the out-migration rate by 1.2 percentage points. The presence of a communist regime also decreases the likelihood of migrants to return to their home country, explaining at least some of the difference between the emigration rates of Mexican and Cuban migrants.

Borjas and Bratsberg also found that emigration is dependent upon the selectivity of immigration. If the immigrant flow was selected on high-skills, the return migrants were more likely to be low-skilled, whereas immigrant flows selected on low-skills were more likely to produce high-skilled out-migrants. Thus, if emigration has a notable impact on the year of entry question, one would expect a low-skilled immigrant flow like Mexicans to experience decreased aggregate educational levels over time, while a high-skilled immigrant flow like Chinese or Asian Indian to experience increased aggregate educational levels over time. It is unclear how substantial an impact these tendencies may create.

Circularity of Migration

Mexican and Central American migrants are the most likely immigrant groups to engage in circular migration—repeatedly migrating to and from the U.S.— and therefore any bias introduced by circularity is most likely to affect them. If the year of entry question measures only the most recent entry, migrants who have made multiple trips (and have accumulated human capital from these trips) will bias upward the estimates of human capital variables for recent arrivals. Lindstrom and Massey (1994) examined this possible bias with data from the Mexican Migration Project (MMP) and found that Mexican migrants accumulate language skills over successive trips to the U.S., so that the duration of the most recent trip has little effect on language acquisition after migrants have accumulated 15 years of experience in the U.S.

Lindstrom and Massey also found that the Census undercounts the most recent immigrants, who are more likely to be temporary and/or undocumented. Thus, while unmeasured experience in the U.S. biases the language abilities of recent Mexican immigrants upward, the undercount of temporary and undocumented migrants biases language abilities downward.

This may be why Espinosa and Massey (1997) found that cross-sectional analysis of census data produced relatively reliable estimates of language acquisition for Mexican migrants. Also, improvements over time were not due to selective emigration, because Mexican migrants who returned to Mexico also had improved English language ability with greater experience in the U.S.

More recently, Redstone and Massey (2003) have challenged the accuracy of the census year of arrival question through their analysis of the New Immigrant Survey pilot study. Findings from a question modeled after the census year of arrival were compared to those that asked about earliest trip to the U.S. and total length of time spent in the U.S. during all trips. (The latter questions were asked in the first wave of the survey, one year before the year of arrival question was asked of the same people.) Redstone and Massey find major inconsistencies between respondents' stated year of arrival and either their date of earliest trip to the U.S. or their date of permanent residency granted by the Immigration and Naturalization Service. They attribute the inconsistency to circular migration, and they show that the estimates of total U.S. experience based on the year of arrival question are biased upward. As a result, the authors show that the effect of experience on earnings is 0.030 using the year of arrival variable but only 0.023 using the total experience variable (although the difference is not statistically significant).

Inconsistency Between the Year of Arrival and Migration Questions

A troubling aspect of the year of arrival data is how well it conforms to data collected by the migration question on the same census questionnaire. In principle, those who came to the U.S. in the interval 1995-2000 should have marked that they were living abroad on April 1, 1995. Although it is possible to have entered the U.S. in the first three months of

1995 and still arrived in the 1995-April 2000 interval, that is a very small portion of the interval (1/21). As discovered by Ellis and Wright (1998), using the 1980 and 1990 data, more than 20% of recent immigrant arrivals say that they were already living in the U.S. five years before the census.

Table 1 summarizes the inconsistency between the two census questions in the 1980, 1990, and 2000 censuses. In 2000, nearly one-third (31.6%) of immigrant respondents who said they arrived in the 1995-2000 interval also said they were residing in the U.S. by April 1 of 1995. This proportion is somewhat higher for Hispanics than for non-Hispanics, and it has gradually increased from 1980 to 1990 and then to 2000.

A more detailed look at the pattern of inconsistency is enabled with 2000 data because Census 2000 asked for exact year of arrival in the U.S. As shown in Figure 1, those who said they came to live in the U.S. in 1995 are clearly most confused about where they were living on April 1 of that year: nearly 70% reported living in the U.S. on April 1, 1995, only one-quarter of the way into the year. This compares to 93% of those who said they arrived a year earlier (1994) and 32% of those who arrived a year later (1996). Nonetheless, even among immigrants who said they came to live in the U.S. in 1999 or 2000, fully 18% claimed that they were living in the U.S. in 1995. Conversely, of those who arrived in 1990 or 1991, as much as 5% said they were not living in the U.S. on April 1, 1995. It is not clear if these discrepancies represent evidence of circular migration or merely respondent error.

Ellis and Wright (1998) theorize about what the inconsistency means in relation to ambiguities and complexities of migration behavior. Primarily, they cast fault on the ambiguity of the meaning of coming to stay in the U.S. First, increasing numbers of immigrants are circular migrants, and the revolving door is not well described by the census question. In addition, many immigrants may not feel that they have come to stay until a date after they actually were first living in the U.S. Apparently, the change in question wording in 2000 (“to live” instead of “come to stay”), and the reordering of questions that brought the migration question immediately after the question on immigration arrival year, have not corrected the inconsistency.

Table 1

**Inconsistency of the Census Questions on Immigrants' Year of Arrival
and Migration Place of Residence Five Years Before the Census**

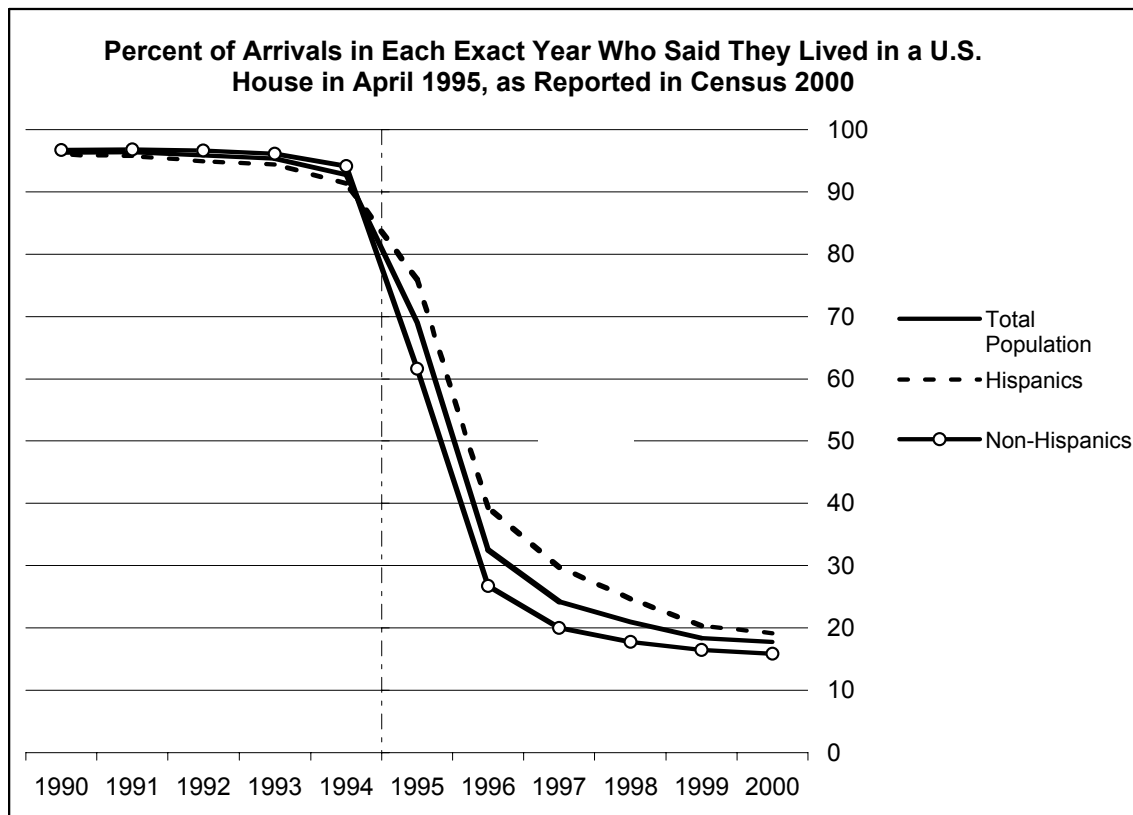
Percent of Recent Arrivals (Within 5 Years of Census Date) Who Reported
Living in the U.S. on April 1 Five Years Before the Census

	<u>1980</u>	<u>1990</u>	<u>2000</u>
Total	25.4	29.5	31.6
Hispanic	34.2	36.4	36.5
Non-Hispanic	20.6	23.7	27.0

Note: Place of residence five years before the census is as of April 1 of 1975, 1985, or 1995 in the respective censuses. Puerto Ricans are excluded.

Source of data: 1980, 1990, and 2000 PUMS 5% files

Figure 1



EVALUATIONS OF DATA QUALITY AND ALTERNATIVES

The foregoing review discloses a series of potential weaknesses in the accuracy of data collected by the census year of arrival question. Let us now turn to a direct examination of the resulting data quality.

Census Bureau Content Reinterview Surveys

The Census Bureau's own evaluations find that the year of arrival question is relatively more reliable than other important questions. Following each decennial census, the Census Bureau conducts an evaluation of data quality. A principal component of this evaluation is a reinterview of a sample of respondents, with the follow up scheduled three to seven months after the April 1 census day. Answers given on the reinterview are compared to those on the submitted census questionnaire to see how consistent are the answers given at different times. Relative stability of responses—i.e., low levels of inconsistency—are interpreted as evidence that the questions provide a reliable measure of the intended concept.

Inconsistency in 2000

The evaluation of Census 2000 found a “low” degree of inconsistency for questions related to place of birth, citizenship and year of entry (Singer and Ennis 2003). Out of 58 items tested, 16 earned the favorable, “low” rating. Other successful items were age, sex, marital status, and veteran status.

In contrast, 16 items were judged to have a “high” degree of inconsistency. These included English-speaking ability, weeks worked last year, amount of income from public assistance, and some questions about disabilities. Even income received from wages earned only a “moderate” rating, as did educational attainment and race. These variables are all staples of socioeconomic analysis and are found to be *less reliable* than the year of arrival variable.

Inconsistency in 1990

The 2000 reinterview study also included historical comparisons. Year of arrival was judged to have a low level of inconsistency in 2000, moderate in 1990, and low in 1980. Review of the reinterview data from the earlier 1990 evaluation study is useful for illustrating how inaccurate are the responses implied by a “moderate” rating (U.S. Census Bureau 1993).

Illustrating this moderate level of inconsistency recorded in 1990, the following percentages of census respondents gave the same response about their year of arrival in the reinterview:

1985-90 on census: 88.4% agreed on reinterview

1980-84 on census: 83.9% agreed on reinterview

1975-79 on census: 76.4% agreed on reinterview

1970-74 on census: 81.4% agreed on reinterview

1965-69 on census: 79.3% agreed on reinterview

1960-64 on census: 81.1% agreed on reinterview

The majority of the inconsistent responses fell into adjacent categories, and if we aggregate to decade instead of 5-year intervals, consistency is markedly improved. For example, of those who indicated on the census questionnaire that they had arrived during the 1980s, 94.0% repeated the same response in the reinterview.

To summarize, the responses to the year of arrival question were judged less accurate in 1990 than in either 1980 or 2000. Nonetheless, even in 1990 there is a fairly high level of consistency. In fact, compared to other 1990 census questions, respondents gave more consistent answers for year of arrival than for factors such as language spoken at home, ancestry or educational attainment.

Implications for Criticisms Based on Circularity

Circularity of migration poses a behavioral challenge to measurement. Yet, it appears that the data inconsistencies which have been interpreted as evidence of circularity may be in part due to respondent error. It is noteworthy that the migration question about place of residence five years ago only earned a “moderate” rating for inconsistency in 2000, a measure of weaker data accuracy than the “low” inconsistency for year of arrival. Thus, the conflict noted by Ellis and Wright (1998) and discussed above may be at least as attributable to respondent error in answers to the place of residence question as the year of arrival question. This is somewhat surprising, because one might presume that the more precise date asked about in the migration question (place of residence on April 1, 1995) is less prone to ambiguity than the general request for year that the person came to live in the U.S.

Similar inconsistencies may well plague the New Immigrant Survey data relied upon by Redstone and Massey (2003) and these have not been evaluated. The evaluation of the year of arrival question in the census is possible because the Census Bureau asks two different questions (immigration and migration) on the same survey instrument and then asks the very same questions of the same respondents seven months later. In either case, there are a noticeable number of respondents who give inconsistent answers. Given that the New Immigrant Survey asks *different* questions at two points in time that are spaced over a *longer* interval, it is not surprising that substantial inconsistencies are reported between the two sets of resulting data. Unfortunately, Redstone and Massey have not

reported the degree of inconsistency that is revealed when the same respondents are asked a second time the same questions, namely what was the date of their first U.S. trip or their total U.S. experience. Would that reinterview reveal as low a degree of inconsistency as the census question on year of arrival, or would it have a moderate degree of inconsistency like the question on place of residence five years ago, or could the reports of total U.S. experience be even more inconsistent? Very likely a sizable amount of inconsistent reporting should be assumed, unless it can be verified otherwise. Therefore, it is not clear how much of this sizable respondent error should be attributed directly to effects of circular migration.

Summary

In 2000, the Census Bureau's reinterview tests showed that responses to year of arrival were remarkably consistent. Respondents generally seem to know what arrival cohort they belong to. Relative to other social and economic questions on the census questionnaire that many analysts have built their careers on, the year of arrival question seems more reliable than many would suspect.

Stability Over Time

Consistency on reinterview is only one measure of data quality. Data could be reliable in that sense but not valid as a representation of experience in the immigrant population over time. The respondents could be consistent in their inaccurate responses, or the sample from which the respondents are drawn could be systematically biased by the emigration and circular migration factors discussed above.

To determine the external validity and stability over time of the data collected by the year of arrival question, we should look at the evidence for a set of cohorts. For this purpose, data are assembled from the censuses of 1980, 1990, and 2000 using the public use microdata 5% file. Data are selected for the entire U.S. (50 states and DC), although all Puerto Ricans are omitted because of their unique circumstances that create confusion surrounding their "arrival" in the U.S. and place of residence. Foreign born residents of the U.S. who reported arrival in 1975-80 and 1970-74 can be observed repeatedly in three successive censuses. These cohorts are segmented into Hispanic and non-Hispanic groupings, and they are restricted to those who were age 20-29 in 1980, 30-39 in 1990 and 40-49 in 2000. Thus the cohorts contain the age group mostly likely to have immigrated and traces the cohort as it advanced into middle age and resides 20 years longer in the U.S.

Key Tests of Stability

With these data we may investigate three questions:

1. How stable is the size of the population in each cohort across the three censuses?
2. How persistent are the size differences between the two arrival cohorts?

3. How stable is the educational attainment level of the cohorts across the three censuses?

In effect, we are testing the consistency over time of the cohorts formed from the year of arrival question. The cohorts are expected to shrink over time through the attrition caused by emigration²; certainly, no new members can join these arrival cohorts after 1980. Attrition also should be greatest in the first decade, and it should be much greater for the newest arrivals, the 1975-79 cohort, than for the 1970-74 cohort. Other than that difference, we would expect the cohorts to maintain consistent size differences over the three censuses.

Also of keen interest is the stability of the educational attainment composition of the cohorts across the three censuses. This is a prime indicator of differential impacts of emigration on the “cohort quality” of the immigrants who have settled in the U.S. Downward sloping educational attainment indicates that the more skilled immigrants have departed, while upward sloping educational attainment indicates that the less skilled immigrants have departed.³ Selective outmigration is one of the greatest threats to the utility of data on duration of residence as collected through the year of arrival variable.

Evidence of Stable Cohort Size

Answers to the above questions can be gleaned from the cohort trajectories displayed in Figure 2. The cohorts generally follow the expected pattern of stable population size over time. As expected, among non-Hispanics there is a sharp attrition for the newest cohort, but that does not appear among Hispanics. Why this fails to appear we will consider below.

The cohorts evidence very slight growth over time. This could reflect respondent error about year of arrival. More likely it reflects differences in census coverage. The greatest likelihood of “undercount” is known to exist for adults in the 20s. These individuals are the most transient in both residence and family status. Recent immigrant arrivals in this age group are surely the most underreported of all population groups in the census (Lindstrom and Massey 1994). Upon maturing into their 30s, a more complete ratio of the cohort is likely to be recorded in subsequent censuses. In addition, the overall

² In this young adult age range we can treat the effects of mortality as negligible.

³ For this purpose we have taken account of the change between 1980 and 1990 in the census questions asked about educational attainment. Whereas the 1980 census asked about years of completed education, the 1990 and 2000 censuses asked about degrees completed. The potential inconsistency of results from the two questions centers on those who have more than a high school degree and less than a four-year college degree. (A series of intermediate technical training options added in 1990 siphoned off persons previously recorded as only high school graduates and inflating the number with “some college.”) Coding was developed to produce comparable response tallies in 1980 and 1990, and this was validated by testing on native-born, non-Hispanic white cohorts passing through the middle age range (whose educational attainment was presumed to be constant from 1980 to 1990).

improvement in census coverage from 1980 to 2000 further accentuates the likelihood of expanded cohort coverage. Thus, the observed size of the arrival cohorts slowly grows over time.

The puzzle is why there is no evidence of initial attrition for the newest Hispanic arrival cohort. A possible explanation must involve the different concept of emigration that attends populations that are engaged in circular migration. Repeated coming and going in this first five years of residence may depress the reported number of “arrived” immigrants when first observed in 1980 (age 20-29). And this circularity may bring back to the sample in 1990 more of the previous arrivals, so that net attrition from emigration is near zero.

Unfortunately, it is not possible to give precise estimates about how many members are lost from each cohort due to emigration, or how many are gained from increased coverage ratios. Therefore, we cannot say exactly how much respondent error may have intruded into the stability of the cohort definition over time. But the net result of these factors appears negligible. The cohorts demonstrate a reasonably stable size across 1980, 1990, and 2000.

Evidence of Stable Cohort Composition

Educational attainment is the single best indicator of cohort skill level over time, because it is the sole measure of individual skills that is stable over time. Occupations, earnings, and English proficiency all grow with experience in the U.S. In contrast, after age 20 high school attainment is fixed, and after age 25 college completion is largely fixed.

What is the net effect on skill composition of churning cohort membership between 1980 and 2000? The evidence in Figure 2 shows that the levels of high school completion and of four-year college completion are basically level and unchanged in all cohorts. The sole exception is with respect to non-Hispanic immigrants that were age 20-29 in 1980. Since about one-third of this cohort may have been too young to have completed college, it would be expected that the proportion rises when they are observed again in 1990 at age 30-39.

On the whole, these data indicate that the arrival cohorts formed from the census year of arrival question are more robust than might be expected. Given the complexities of migration behavior and concerns about respondent accuracy, some might assume that the data are so fraught with error as to be severely biased. The evidence in Figure 2, however, suggests that data users can have confidence in the meaningfulness of analysis with the data from the year of arrival question.

DISCUSSION: APPROPRIATE CAUTION IN USE

New immigrants to the U.S. are surely among the most difficult subjects to accurately sample and from which to collect data. Highly mobile and transient in their initial years in the U.S., and in many cases undocumented in their residency, immigrants prove very difficult subjects for data collection. All efforts to survey this population yield flawed results. Even the carefully crafted New Immigrant Survey pilot study was able to retrieve data from only 61% of legal immigrants admitted by the INS (Jasso et al., 2000). The relevant question is not whether the data are perfectly accurate; rather the appropriate questions are whether the data from a given survey are so biased as to be unusable or under what conditions we can maximize the accuracy of inferences drawn from the data.

Appropriate caution should be taken when using data collected by the census question on immigrants' year of arrival. The evidence discussed above shows that by far the lowest reliability of data pertains to immigrants with less than five years residence in the U.S. It is these short-term residents who are most confused about their year of arrival. Perhaps the confusion stems from practices of circular migration, or it may derive from fundamental ambiguity about the meaning of "come to stay [or live]." Analysts should treat these short-term residents with special care.

A design similar to Chiswick's (1978) original cross-sectional formulation of Years Since Migration was adopted by Redstone and Massey (2003) in order to show how inaccuracies of estimated years of U.S. experience bias upward the experience-earnings elasticity. This formulation is highly vulnerable to error because it incorporates data for the least reliable, short-term immigrants.

An alternate approach introduced by Borjas (1985), maintained the pooling of short and long term immigrants, but separately identified cohorts with dummy variables. The same duration function of experience-earnings elasticity was imposed, save that specific cohorts were shifted to higher or lower parallel trajectories. This formulation is improved over the simple cross-sectional design, but the duration function is still vulnerable to error concentrated in the short-term immigrants.

Yet a third approach has been employed less frequently (e.g. Myers and Cranford 1998, Duleep and Dowhan 2002). Each cohort is not only uniquely identified by a dummy variable, but a unique duration effect on the outcome of interest is estimated for each cohort. By not imposing a time path from one cohort on others, this has the advantage of containing the error within the shorter-term cohort. It also allows one to identify unique behaviors (and data problems) that may pertain to one or more cohorts. In general, it is always desirable to separately examine the time trends for each cohort rather than blindly pool all the data in a single estimation.

An additional consideration is to pool immigrants in arrival cohorts defined by wider time spans (i.e., 10 years instead of 5), obviating some of the documented inconsistency problems found in census reinterview studies. Also, studies should always incorporate educational attainment as a control to protect against bias from shifts in cohort

composition over time, even though the present study has not found evidence of such bias.

CONCLUSION

This working paper has reviewed the rationales expressed for doubting the accuracy of data collected by the census year of arrival question asked of the foreign born. There is good theoretical basis for wondering if the data are sufficiently reliable to support analysis. However, empirical analysis testing the reliability and validity of the year of arrival data provides substantial assurance of relative data accuracy.

The conclusion reached here is that the data are much more reliable than some may have feared. Census reinterview studies show that respondents provide more consistent answers to the year of arrival question than to many other important social and economic variables, including race and educational attainment. In addition, inspection of cohorts from 1980 to 2000 also provides reassurance of data validity. Longitudinal trends in the size of arrival cohorts and in their educational composition indicate great stability.

The only exceptional cohort is comprised of those that arrived in the last five years preceding the census, a group well known to be transient and under-reported. The non-Hispanic portion of this cohort sustained substantial attrition from emigration, while the Hispanic portion did not. A tentative explanation for this puzzle is that this groups was more heavily exposed to circular migration that blunted the effects of attrition that were expected to be recorded.

These exceptions are consistent with those feared by critics of the year of arrival data. However, the problems appear to be relatively confined to short-term immigrant arrivals. Accordingly, we have discussed some general cautions and procedures that should be observed in order to enhance the quality of inferences to be drawn from the year of arrival data.

Overall, the value of the census or CPS year of arrival data must be weighed against the alternatives that are feasibly available. First, there is no known data set that can reconstruct an historical record that is comparable to that afforded by the census year of arrival data since 1970. Surely this question must be preserved in the future if we are to be able to compare the experiences of different immigrant arrival waves over time.

Second, despite problems of census coverage and respondent error, there is no other source of data that can supply comprehensive information about the full extent of the foreign born population, including not only legal but undocumented residents. More research is needed to learn how we might supplement or adjust these data to provide the most accurate picture. But surely the census year of arrival data will retain a central role in any data retrofitting.

Finally, the census year of arrival data remain an indispensable means of estimating longitudinal progress for immigrants in the U.S. The consistency of the data over time (at least after the first five years) and the large sample sizes make the decennial census

data, and to less extent the Current Population Survey, invaluable for tracing net changes in each wave of immigrant arrivals as they settle longer, and assimilate, over the decades.

REFERENCES

- Ahmed, Bashir and J. Gregory Robinson (1994). "Estimates of emigration of the foreign-born population: 1980-1990". U.S. Bureau of the Census, Population Division Working Paper No. 9, Washington, DC.
- Alba, Richard D. and Victor Nee (2003). *Remaking the American mainstream: Assimilation and contemporary immigration*. Cambridge, MA: Harvard University Press.
- Borjas, G. J. (1985). "Assimilation, changes in cohort quality, and the earnings of immigrants," *Journal of Labor Economics*, 3:463-489.
- Borjas, George J and Bernt Bratsberg (1996). "Who leaves? The outmigration of the foreign-born," *The Review of Economics and Statistics* 78:165-176.
- Chiswick, Barry (1978). "The effect of Americanization on the earnings of foreign-born men," *Journal of Political Economy* 86:897-921.
- Costanzo, Joseph M., Cynthia J. Davis, and Nolan Malone (2002) "Guide to International Migration Statistics: The Sources, Collection and Processing of Foreign Born Population Data at the U.S. Census Bureau," Population Division Working Paper No. 68.
- Duleep, Harriet Orcutt and Mark C. Regets (1997). "The decline in immigrant entry earnings: Less transferable skills or lower ability?" *Quarterly Review of Economics and Finance* 37:189-208.
- Duleep, Harriet Orcutt and Dan Dowhan (2002). "Insights from Longitudinal Data on the Earnings Growth of U.S. Foreign-Born Men," *Demography* (August): 485-506.
- Ellis, Mark and Richard Wright (1998). "When immigrants are not migrants: Counting arrivals of the foreign born using the US census," *International Migration Review* 32:127-144.
- Espinosa, Kristin E. and Douglas S. Massey (1997). "Determinants of English proficiency among Mexican migrants to the United States," *International Migration Review* 31:28-50.
- Gibson, Campbell J. and Emily Lennon. 1999. "Historical census statistics on the foreign-born population of the United States: 1850-1990". U.S. Census Bureau, Population Division Working Paper No. 29, Washington, DC.
- Jasso, Guillermina, Douglas S. Massey, Mark R. Rosenzweig, and James P. Smith. 2000. "Data sources and data quality - The New Immigrant Survey Pilot (NIS-P):

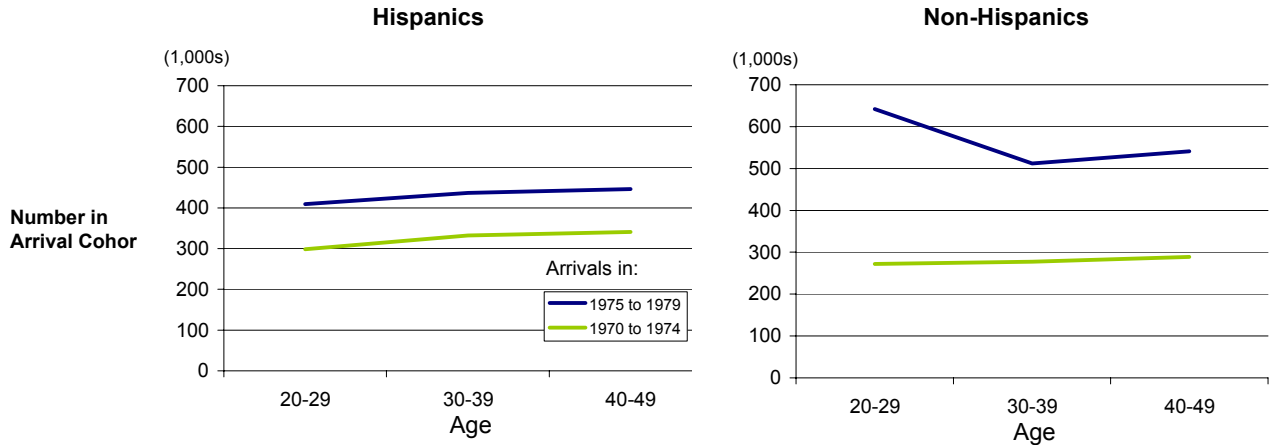
Overview and new findings about U.S. legal immigrants at admission."
Demography 37(1):127-138.

- Lindstrom, David P. and Douglas S. Massey (1994). "Selective emigration, cohort quality, and models of immigrant assimilation," *Social Science Research* 23:315-349.
- Myers, D., & Cranford, C. (1998). "Temporal differentiation in the occupational mobility of immigrant and native-born Latina workers," *American Sociological Review*, 63:68-93.
- Myers, D., & Lee, S. W. (1996). "Immigration cohorts and residential overcrowding in Southern California," *Demography*, 33:51-65.
- Portes, Alejandro and Min Zhou (1993). "The new second generation: Segmented assimilation and its variants," *Annals of the American Academy of Political and Social Science* 530:74-96.
- Redstone, Ilana and Douglas S. Massey (2003). "Coming to Stay: An Analysis of the U.S. Census Question on Year of Arrival," Paper presented at the PAA meetings in Minneapolis.
- Singer, Phyllis and Sharon R. Ennis (2003). "Census 2000 Content Reinterview Survey: Accuracy of Data for Selected Population and Housing Characteristics as Measured by Reinterview". Census 2000 Evaluation B.5.
- U.S. Census Bureau (2002). "Measuring America: The Decennial Censuses from 1790 to 2000". POL/02-MA.
- U.S. Census Bureau (1993) "Content Reinterview Survey: Accuracy of Data for Selected Population and Housing Characteristics as Measured by Reinterview". 1990 CPH-E-1.

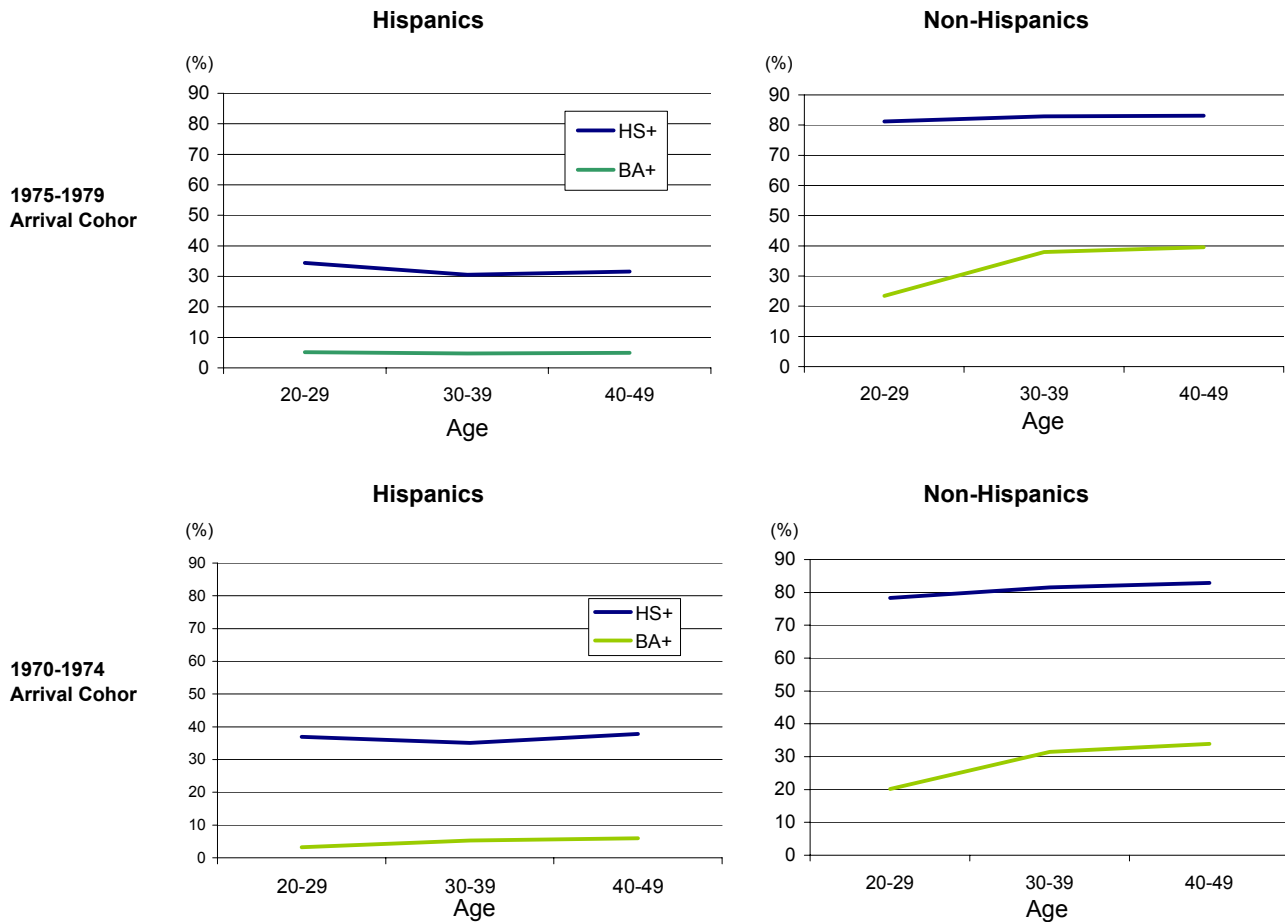
Figure 2

Stability of Arrival Cohorts Observed Repeatedly in 1980, 1990 & 2000

Population Size



Educational Attainment



Source: 1980, 1990, 2000 5% PUMS Data