How Well Would Administrative Records Correct the Undercount of Young Children in the U.S. Census?

By

Dr. William P. O'Hare

Over the past couple of decades, the Census Bureau has been steadily building a robust administrative records (AR) database.[1] The most recent version of the Census Bureau AR database combines data from 31 different sources (see page 45 of U.S. Census Bureau 2023, for a list of these sources)

The Census Bureau used this AR database to conduct what it calls a "2020 AR Census", essentially simulating the 2020 decennial census using these administrative records as opposed to asking people for self-response and then following up with people who don't respond. A recent Census Bureau report (2023) compares demographic characteristics in the AR census to the demographic characteristics of people in the 2020 Census.

This new Census Bureau report is important because administrative records are increasingly being discussed as a way to improve future U.S. censuses. Therefore, it is important to see how well administrative records reflect different populations by examining differences in demographic characteristics between an AR census and the 2020 census count.

This paper draws on some data from that Census Bureau report to examine how the data from the AR census compares to the data from the 2020 census for young

[1] Administrative records data is generally used to refer to data that are collected for administrative purposes, not statistical purposes. Birth and death certificate data are a good example.

children, defined here as ages 0 to 5 (in Census Bureau jargon, children under age 1 are referred to as age “0”). The focus is on young children because that age group had the highest net undercount of any age group in the 2020 Census based on the Census Bureau’s Demographic Analysis. Also, better use of administrative records has been one proposal for reducing the undercount of young children in the 2030 Census.

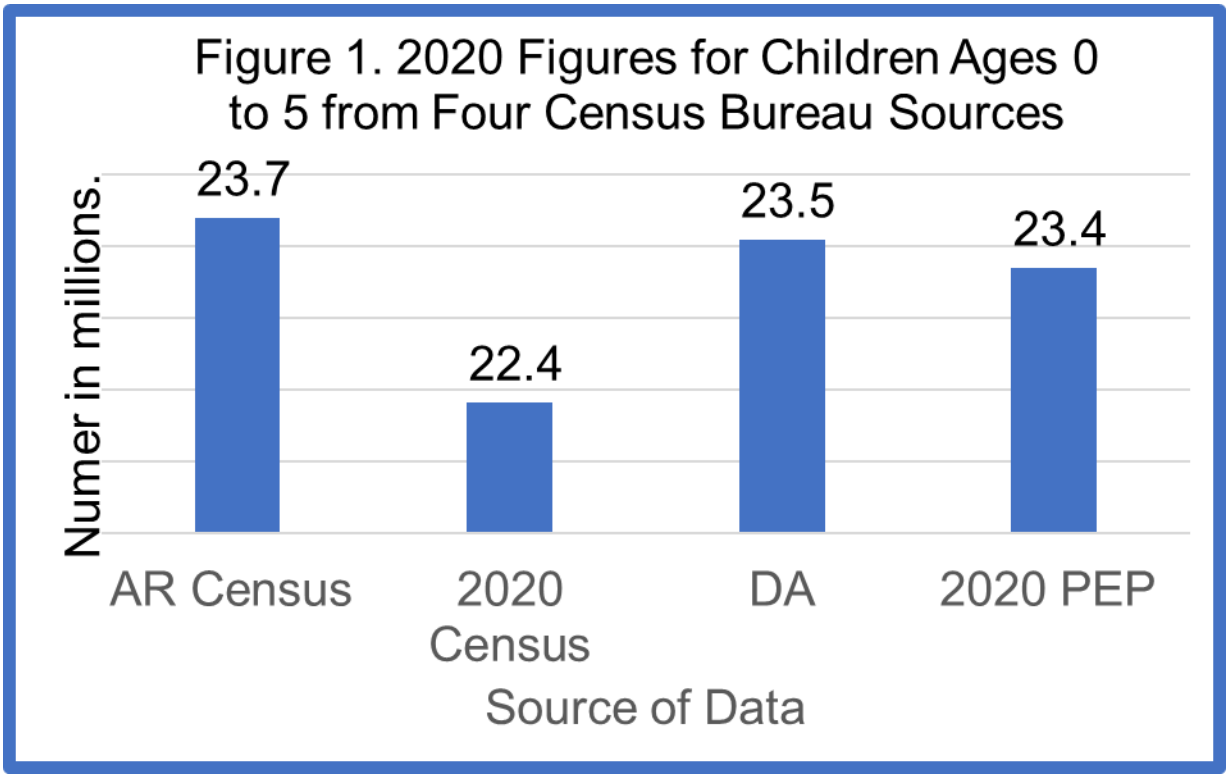
This Census Bureau report helps answer the question about what the 2020 decennial census would have looked like if it had been based entirely on administrative records. This Census Bureau report also addresses how well administrative records reflect data for young children.

How Does Data from the AR Census Compare to the Count of Children Ages 0 to 5 from Different Census Bureau Sources?

Figure 1 shows the number of children ages 0 to 5 from the 2020 AR census, the 2020 census, the Census Bureau’s Demographic Analysis (DA), and the vintage 2020 Population Estimates Program (PEP). DA estimates are usually seen as the most accurate source of data for young children, in part because they are largely based on birth certificate data and that data source is very reliable.

There are a couple of implications from the data in Figure 1. First, examining all four series together it is clear the 2020 census count is the outlier. For ages 0 to 5, the total count of young children from the AR census, DA, and the 2020 PEP are all well over 23 million while the figure from the 2020 census is 22.4 million. All these comparisons suggest a 2020 census undercount of young children in the neighborhood of one million or more. This underscores the extent to which the 2020 census figure for the young child population is a serious undercount based on comparisons with multiple

sources of data. Also, the AR census figure is similar to DA estimates for ages 0 to 5, which suggests that the AR figure is credible and is probably a more accurate count of young children than the Census count.



How Does Data for Ages 0 to 5 from the AR Census Compare to the 2020 Census?

Table 1 shows the population figures from the AR census and the 2020 census for ages groups 0 to 2, 3 to 5, and collectively for both age groups combined. The AR census shows 23.7 million children ages 0 to 5 compared to 22.4 million for the 2020 Census. This means for the population ages 0 to 5, the decennial census count for young children was about 1.3 million children below the AR census and this amounts to a 5.4 percent difference.

Comparison of 2020 census and Demographic Analysis estimates for the population ages 0 to 5, shown in the next section of this paper, shows the population ages 0 to 5 had a 4.8 percent net undercount in the 2020 census.[2] In other words, if the 2020 Census had been based totally on AR, there would not have been an undercount of young children. This demonstrates the promise of using AR in the future.

Table 1 shows the results for ages 0 to 2 and ages 3 to 5 are similar to each other in terms of the relationship between the AR census and the 2020 census. The difference between the number of children ages 0 to 2 is 5.5 percent and the difference for children ages 3 to 5 is 5.4 percent.

Table 1 Comparison of Young Child Populations from AR Census and 2020 Census

	numbers in 1000s			
	AR census	2020 Census	Numeric Difference (2020 Census - AR Census)	Percent Difference ((2020 Census-2020 Census)/AR Census) *100
ages 0 to 2	11,310	10,690	-620	-5.5
ages 3 to 5	12,390	11,720	-670	-5.4
ages 0 to 5	23,700	22,410	-1,290	-5.4
Source: U.S. Census Bureau, 2023, Table 75, page 178				
Note Undercounts are shown as negative numbers				

[2] Note the most common undercount rate reported for young children is for age 0 to 4, where the net undercount rate in the 2020 census was 5.4 percent. The Census Bureau report I am discussing here provides data for ages 0 to 5, not 0 to 4.

How Does the AR Census for Young Children Compare to Demographic Analysis?

Another way of looking at the AR census and the 2020 census for young children is to compare them to the DA estimates. As stated earlier in this paper, DA is considered the gold standard for the number of young children. Data displaying comparisons of the AR Census, the official 2020 census, and DA are shown in Table 2.

Table 2. Comparison of Young Child Populations form AR Census, 2020 Census to Demographic Analysis

	numbers in 1000s			numeric difference between 2020 Census and DA (2020 Census - DA) (in 1,000s)	percent difference between 2020 Census and DA (2020 Census - DA)/DA)*100	numeric difference between AR Census and DA (AR Census - DA) (in 1,000s)	percent difference between AR Census and DA (AR Census - DA)/DA)*100)
	AR census	2020 Census	DA*				
age 0-2	11,310	10,690	11,418	-728	-6.4	-108	-0.9
age 3-5	12,390	11,720	12,130	-410	-3.4	260	2.1
age 0-5	23,700	22,410	23,548	-1,138	-4.8	152	0.6
Source: U.S. Census Bureau, 2023, Table 75, page 178							
Note undecounts are shown as negative numbers							
*middle series DA							

Comparisons to Demographic Analysis estimates shows the population ages 0 to 5 had a 4.8 percent net undercount in the 2020 census while the AR census had a slight (0.6%) overcount. The AR census seems to more than make up for census undercounts of young children in the 2020 census.

For both age groups the AR census is much closer to the DA figure than the 2020 census but the two age groups differ in the relationship between the AR census and the 2020 census. Based on comparisons to DA, the net undercount for age 0 to 2 in the 2020 census was 6.4 percent. On the other hand, the net undercount of age 0 to 2 based on the AR census is only 0.9 percent. Thus, the AR census seems to largely eliminate the undercount in the 2020 census for children age 0 to 2.

Based on DA, the net undercount for ages 3 to 5 in the 2020 census was 3.4 percent. But there was a net overcount of 2.1 percent for ages 3 to 5, based on comparing the AR census to DA. Thus. the AR census seems to more than make up for the undercount in the 2020 Census for children age 3 to 5 and actually produces and overcount for this group.

This analysis shows there are a larger share of children ages 3 to 5 captured in AR than for children ages 0 to 2. It is not clear to me why this age difference occurs? The Census Bureau should explore why the age discrepancy occurs and look for additional sources of administrative records that better capture the youngest children.

How Well Does the AR Census Reflect Young Hispanic?

For decennial census figures, the overall accuracy is not as important as differential accuracy. Differential accuracy (often presented as differential census undercounts) across locations and demographic subgroups can lead to inequities in the distribution of resources.

In that context, the overall difference based on comparing AR census to the 2020 census for the population ages 0 to 5 masks important variation in quality. In short, the AR data for all young children is promising but the situation for young Hispanic children is problematic.

I present data for young Hispanic children here, but I do not present equivalent data for young Black children because I don't think this is a credible comparison. For all Black children age 0 to 5, the AR census is 23 percent higher than the 2020 Census. I have to believe this difference has something to do with how race is collected and

coded in these two data sources. For example, the Census Bureau will be releasing a Modified Race File soon that will contain adjusted numbers of Black children reported in the Census by moving those individuals who marked “Some Other Race” into one of the five major race categories. In addition, I am sure there are a lot of complexities related to how data on race is collected and reported in the 31 administrative records sources used by the Census Bureau.

How Well Does the AR Census Reflect Young Hispanic Children?

Young Hispanic children had high net undercounts in the 2020 Census revealed a comparison of the 2020 census count to DA estimates. There was a 7.9 percent undercount of Hispanic children age 0 to 5 based on DA. This is similar to the 2010 Census where Hispanics ages 0 to 4 had an official undercount of 7.5 percent.

Detailed data for young Hispanic children are shown in Table 3. For ages 0 to 5, there is little difference between the AR census and the 2020 census for the number of young Hispanic children. The AR census reports 5,630,000 young Hispanic children compared to 5,650.000 from the 2020 Census. That means 20,000 fewer children in the AR census than in the 2020 census and that amounts to a 0.4 percent difference. In other words, unlike the data for all children, an AR census would not eliminate the undercount of Hispanic children ages 0 to 5 because the AR census and the 2020 census results are almost identical.

Table 3. Comparison of Young Hispanic Child Populations form AR Census and 2020 Census				
	numbers in 1000s		numbers in 1000s	
	AR census	2020 Census	Numeric Difference (2020 Census- AR Census - 2020 Census	Percent Difference (2020 Census -AR Census/AR Census) *100
ages 0 to 2				
Male	1,366	1,366	0	0.0
Female	1,412	1,319	-93	-6.6
Total ages 0 to 2	2,778	2,685	-93	-3.3
ages 3 to 5				
Male	1,472	1,508	36	2.4
Female	1,380	1,457	77	5.6
Total ages 3 to 5	2,852	2,965	113	4.0
ages 0 to 5				
Male	2,778	2,685	-93	-3.3
Female	2,852	2,965	113	4.0
Total ages 0 to 5	5,630	5,650	20	0.4
Source: U.S. Census Bureau, 2023, Table 75, page 178				
Note Undercounts are shown as negative numbers.				

Data in Table 3 are broken down by age and sex because there are some strange patterns for young Hispanic children based on these demographic characteristics.

For Hispanic children ages 0 to 2 the AR census is higher than the 2020 census, but for Hispanic ages 3 to 5, the 2020 census is higher than the AR census. For ages 0 to 2, the 2020 census count of young Hispanic children is 3.3 percent lower than the AR census. On the other hand, for the age group 3 to 5, the AR census is 4.0 percent higher than the 2020 census. This suggests that the youngest children (ages 0 to 2) are over-represented in the administrative records used in the AR census compared to the 2020 census, but for ages 3 to 5 they are under-represented in AR compared to the 2020 census. It is not clear to me why this age pattern exists. The Census Bureau

should look more closely at the data for young Hispanic children to uncover the reason(s) for this unusual pattern.

There is also an interesting pattern by sex for young Hispanic children. For children in these age groups, one would expect the number of males to be very similar to the number of females. If anything, one would expect more males than females because the sex ratio at birth in the U.S. is about 105, meaning there are 105 males born for every 100 females.

But the data for Hispanic children ages 0 to 2 shows the females outnumber males in the AR census by 46,000 or 3 percent (1,366,000 males to 1,412,000 females). On the other hand, for ages 3 to 5, the reverse is true; males outnumber females by 92,000 or 6 percent (1,472,000 males compared to 1,380,000 females) in the AR database. For males, the AR data are 2.4 percent below the 2020 census count, and for females the AR data are 5.6 percent below the 2020 Census data.

It is not immediately apparent to me why the number of males and females from the AR census have this peculiarity, but it raises questions about the quality and completeness of AR data for young Hispanics. The Census Bureau should investigate these strange patterns among young Hispanic children.

There are a couple of issues addressed in the Census Bureau report that might have special implications for data on Hispanics. The report authors note that non-citizens are likely to be under-represented in administrative data, and they also note the census counts in many of the counties on the Texas/Mexican border are problematic.

Do Administrative Records Have Good Address Data?

While the comparisons of national totals from the AR census and the 2020 census show administrative records to be a promising approach to improving the count of young children, it is important to remember that the use of administrative records in the census is more complicated than just comparing national totals from administrative records for different demographic groups. To be useful in a census, we must know what housing unit a person in an administrative record is attached to. The data provided in the Census Bureau report indicate there are a lot of administrative records that could not be connected to the appropriate housing unit.

Summary

While the information reported here is promising, further work needs to be done. First it is important to understand how well young children reflected in administrative records can be placed into the correct housing unit and what can be done to improve that. It is also important to investigate why young Hispanic children seem to be under-reflected in administrative records and why there is a difference between the youngest Hispanic children (ages 0 to 2) and Hispanic children ages 3 to 5. Further investigation is needed to understand why the ratio of young Hispanic males and females are different by age.

The recent Census Bureau report analyzed here shows that collectively the AR census provides a more accurate figure for young children than the 2020 census count. An AR Census in 2020 would have eliminated the 2020 Census undercount for young children at the national level. This indicates further use of AR may help reduce or eliminate the chronic undercount of young children in the U.S. Census.

However, the figure for all children masks some differences in quality for subgroups. It appears that children ages 0 to 2 are under-represented in the AR database, while children ages 3 to 5 are over-represented. Finding more administrative records that reflect the youngest children is important.

For young Hispanic children, an AR census would do little to correct the undercount in the 2020 census. Given the results for young Hispanic children it would be wise for the Census Bureau to investigate how data on race and Hispanic Origin are collected and coded in administrative records compared to the Census. For young Hispanic children, the AR census shows an undercount similar to the 2020 Census.

In addition, for purposes such as political redistricting and allocating federal funds, an AR census would be seriously flawed because of the inability to link data for many individuals to specific unique households.

References

U.S. Census Bureau (2023). "Real-Time 2020 Administrative Record Census Simulation," Version 1.0 issued May 5, 2023. J. David Brown, Samuel R. Cohen, Genevieve Denoeux, Suzanne Dorinski, Misty L. Heggeness, Carl Lieberman, Linden Mc Bride, Marta Murray-Close, Hangxun Qin, Allen E. Ross, Danielle H. Sandler, Lawrence Warren, and Moises Yi. https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/evaluate/eae/2020-admin-record-census-simulation.html