Addressing Data Collection Errors in the Fertility Question in the American Community Survey

By Tavia Simmons

SEHSD Working Paper 2016-21

August 2016

In recent years, a few geographic areas in the American Community Survey (ACS) data had unusually high percentages of women reported as giving birth in the past year, quite unlike what was seen in previous years for those areas. This paper describes the issue that was discovered, and the measures taken to address it.

ACS data collected in Williamson County, TX illustrate the problem. In 2012, 12.2 percent of women aged 15-50 in Williamson County, TX were reported as having given birth in the past year (see Table 1). This was in stark contrast to previous years: 7.0 percent in 2009, 6.4 percent in 2010, and 6.4 percent in 2011.[1] Considering the large increase from past years' results, we investigated further.

Table 1- Percent of women aged 15 to 50 who gave birth in the last year in Williamson County, TX

Year	Percent	MOE[1]
2009	7.0	1.6
2010	6.4	1.3
2011	6.4	1.8
2012	12.2	2.1

[1.] MOE is short for 'margins of error'. Data are based on a sample and are subject to sampling variability. A margin of error is a measure of an estimate's variability. The larger the margin of error in relation to the size of the estimate, the less reliable the estimate. When added to and subtracted from the estimate, the margin of error forms the 90 percent confidence interval.

Source: U.S. Census Bureau, American Community Survey, 2009 through 2012 1-year data

ACS data were collected via mail, phone interviews, and personal interviews.[2] When we looked at the data by interview mode, it was clear that a higher percentage of the cases in which women were reported as having given birth were collected via personal interviews in 2012 than in recent years. While from 2009 to 2011, 38.5 to 50.5 percent of data on women in households with a recent birth collected in

[1] There was no significant difference between the 2009, 2010, and 2011 percentages of women who gave birth in the last year in Williamson County, TX.

[2] Starting in 2013, ACS data was also collected via the internet.

Williamson County came from Computer Assisted Personal Interviews (CAPI), in 2012, 71.1 percent of these data came from CAPI interviews (see Table 2).[3]

Table 2- Distribution of women aged 15 to 50 who gave birth in the last year, by mode, Williamson County, TX (For women in households)

Year	Total	MOE[1]	Mail	MOE	CATI	MOE	CAPI	MOE
2009	100.0	x	44.0	12.6	14.5	5.8	41.4	13.4
2010	100.0	x	58.8	11.0	2.7	2.1	38.5	10.7
2011	100.0	x	34.2	12.6	15.3	7.8	50.5	13.9
2012	100.0	x	25.0	6.7	3.9	2.2	71.1	7.4

X Not applicable.

[1] MOE is short for 'margins of error'. Data are based on a sample and are subject to sampling variability. A margin of error is a measure of an estimate's variability. The larger the margin of error in relation to the size of the estimate, the less reliable the estimate. When added to and subtracted from the estimate, the margin of error forms the 90 percent confidence interval.

Source: U.S. Census Bureau, American Community Survey, 2009 through 2012 1-year data

Given the apparent increase in women giving birth originating from CAPI data, we took a closer look at the data coming in from individual field representatives (FRs) working in that geographic area. One particular FR had a strikingly high reported percentage of women 15-50 years old who had given birth in the last year (see Table 3). For this FR's interviews, 62.9 percent of all women 15 to 50 years old were reported to have given birth in the last year, compared to 7.6 percent of women 15 to 50 years old from interviews conducted by other FRs active in the same county.

Table 3- Percent of women aged 15 to 50 who gave birth, for CAPI* responses in Williamson County, TX, 2012 ACS (For women in households)

FR type	Percent	MOE[1]
flagged FR	62.9	11.5
All other FRs	7.6	3.1

* CAPI stands for Computer Assisted Personal Interview

[1] MOE is short for 'margins of error'. Data are based on a sample and are subject to sampling variability. A margin of error is a measure of an estimate's variability. The larger the margin of error in relation to the size of the estimate, the less reliable the estimate. When added to and subtracted from the estimate, the margin of error forms the 90 percent confidence interval.

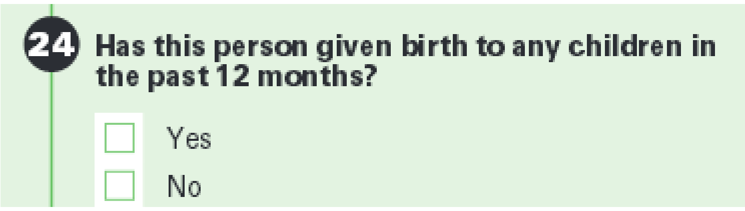
Source: U.S. Census Bureau, American Community Survey, 2012 1-year data

During the latter part of 2013, we found a handful of cases across the country where similar results to those seen in Williamson County, TX were found. In most cases, FRs admitted when asked that they may not have always asked the full question. Figure 1 provides an image of the fertility question. Since the

[3] There was no significant difference between the 2009, 2010, and 2011 percentages of women who gave birth in the last year in Williamson County, TX by CAPI interview.

phrase “in the past 12 months” falls at the end, if the full question is not read, the question takes on a vastly different meaning—“Have you given birth?” versus “Have you given birth *in the last 12 months*?” This could easily lead to the much higher percentages of women reported as having given birth for these geographies.

Figure 1. American Community Survey fertility question



24 Has this person given birth to any children in the past 12 months?

□ Yes
□ No

Given the implausibly high fertility rates discovered for these scattered geographies, we needed to address the issue in the publically released data. Our initial solution in the latter part of 2013 was a technique called geography suppression—where estimates for the affected geographies were not published. While this effectively eliminated the release of low quality data, there were a number of downsides. First, it was incredibly labor intensive, involving identifying which geographies were affected enough to require suppression as well as many other steps within the automated system to ensure that affected estimates were not released. Second, the time required to complete these extra tasks was difficult to fit into the regular processing schedule. Third, and most importantly, no fertility estimates were shown for the affected geographies. For smaller geographic areas that are only covered in the 5-year ACS file, this meant they received no estimates for 5 years.

Given the disadvantages of suppressing low quality estimates, we began to develop alternative ideas to address the issue. We implemented annual training to remind FRs to ask the full question, even if the respondent cuts them off before they finish. The training explained how omitting those last five words could produce radically different estimates of women who gave birth in the last year.

The more direct solution we developed to replace suppression involved creating a system of carefully crafted rules that would allow us to flag FRs with unusually high fertility estimates. One major benefit to this method is that the data are reviewed soon after data collection, before it is processed, rather than nearly a year later, during review of the data, after processing and editing. The rules had to be carefully calibrated to avoid false positives --such as geographies with legitimately high percentages of births. For example, Kiryas Joel, New York, has estimates of 26.2 percent of women 15 to 50 years old having given birth in the last 12 months.[4] This is a reasonable estimate, however, since Hasidic Jewish communities often have high levels of fertility. At the same time, the rules must catch those geographies with suspiciously high percentages of women giving birth even though the demographic characteristics of the

[4] These estimates are from the 2010-2014 American Community Survey 5-year data.

areas may vary widely. The rules need to catch data collection issues in places like college towns with higher proportions of younger women who have never given birth, as well as areas with an older population where fewer women have given birth recently, but the majority are mothers.

In creating the conditions under which problem FRs would be flagged, we focused on a combination of 1) areas where FRs were reporting abnormally high percentages of women giving birth in the last year and 2) there were low percentages of infants present in their households. If cases for a particular FR meet the conditions for that data year, their cases with a 'yes' value for the fertility question will be blanked out and allocated during the editing procedures.

The exact rules are as follows:

Conditions under which to **blank out cases and allocate**:

1. If this FR has at least 30 cases of women aged 15-50 who were interviewed AND
2. At least 30 percent of those women aged 15-50 interviewed by this FR have a value of 'yes' for whether they had a birth in the past year AND
3. 75 percent or fewer of those cases where women were reported to have had a recent birth have a child age 0 or 1 in the household,

THEN blank cases for that FR where women were reported to have had a recent birth.

Additionally, an 'early warning system' was also added for less extreme cases, where FRs would be contacted, but the data would not be blanked and allocated:

Contact FRs who:
1. have interviewed at least 100 women aged 15-50 AND
2. have at least 15 cases where women were reported to have had a recent birth AND
3. 75 percent or fewer of the women reported to have had a recent birth have a child age 0 or 1 present in the household.

We implemented this allocation system during data collection in 2013. We take a look at the effect of using these rules on 2013 data, focusing on a few geographies that had data suppressed in 2012 (see Table 4).[5] In these geographies, the percentage of women who gave birth in the last year was starkly different in 2012 versus 2013. For instance, in Floyd County, Georgia, the percentage who gave birth in the last year dropped from 32.8 percent in 2012 to 4.1 percent in 2013. In Durham County, NC, a locale with a relatively younger population due to several colleges and universities located there, the percentage who gave birth in the last year dropped from 12.2 percent in 2012 to 3.4 percent in 2013.[6] As expected, in concert with the drop in the percentage who gave birth in the last year, the allocation rate for fertility data for these geographies increased from 2012 to 2013. For example in Floyd County, Georgia, the allocation rate for the fertility question went from 4.0 percent in 2012 to 24.5 percent in 2013.

[5] Margins of error for Table 4 are presented in Appendix Table 1 at the end of this document.

[6] There was no significant difference between the percentages of women who gave birth in the last year in Floyd County, Georgia and Durham County, North Carolina in 2013.

Table 4: Effect of 'blanking and allocating' method in 2013 on geographies flagged in 2012 as having a suspiciously high percentage of women aged 15-50 who gave birth in the last year

	Nation	Bartow County, GA	Cherokee County, GA	Floyd County, GA	Durham County, NC	Franklin County, OH	Dallas County, TX	Williamson County, TX
Percent of women who gave birth, 2013	5.2	2.8	4.5	4.1	3.4	5.5	5.2	4.9
Percent of women who gave birth, 2012	5.4	24.7	16.3	32.8	12.2	7.7	8.2	12.2
Percent of fertility data allocated, 2013	6.7	12.6	17.1	24.5	7.9	7.2	6.6	10.1
Percent of fertility data allocated, 2012	4.0	2.1	4.5	4.0	2.7	2.6	2.3	3.8
Percent of women who gave birth--for flagged FR, 2013	X	5.0	4.2	3.9	--	1.4	4.6	3.7
Percent of women who gave birth--for flagged FR, 2012	X	63.0	79.8	67.4	48.8	44.3	64.8	62.9
Percent of fertility data allocated--for flagged FR, 2013	X	29.9	38.0	43.6	59.6	37.9	32.6	33.0
Percent of fertility data allocated--for flagged FR, 2012	X	1.3	1.7	--	4.3	0.4	0.8	--

-- Represents that the estimate is zero or rounds to zero

X While overall national level data are included for comparison purposes, flagged FR data was not compiled at the national level since the focus of this table was the effect of individual FRs on specific smaller geographies.

Source: American Community Survey, 2012 and 2013 1-year data -

Focusing on data collected by FRs flagged in 2012, the change from 2012 to 2013 in the estimate of women who had given birth in the last year was even more extreme. For example, in Cherokee County, Georgia, in 2012, based on the flagged FR's returns, 79.8 percent of women 15 to 50 years old were reported to have given birth in the last year. In 2013, after this FR's data had been blanked and allocated, only 4.2 percent were estimated as having given birth recently. Allocation rates for the fertility data collected by these flagged FRs also rose significantly, as expected, from 2012 to 2013. For the data collected by the flagged FR in Durham County, North Carolina, 4.3 percent was allocated in 2012 while 59.6 percent was allocated in 2013, for example.

Overall, ten FRs in data year 2013 had their data blanked and allocated as a result of the new system (including several of the FRs who were discovered in late 2013 during review of 2012 data). In the following year (2014), only 3 FRs were contacted and had their data blanked and allocated—a significant improvement over the 10 contacted the previous year, and they were contacted several months sooner, for corrective measures, compared to the previous year. Based on the drop in flagged FRs in 2014, it appears the annual training we implemented had a positive effect as well.

Having analyzed the results of this method of blanking and allocating data, it appears that this method is superior to the geography suppression method in a variety of ways. It is less labor intensive than geography suppression and it flags FR issues far earlier in the data collection process. In addition, the FR is contacted and the issue dealt with earlier in the process. Unlike geography suppression, this method also allows the fertility estimates to be published for these geographies. One downside to this method is that it results in much higher fertility allocation rates for these specific geographies. However, considering the many benefits to using this method, this is our preferred method. The timeliness of contacting FRs in the field in the middle of the data collection year is a particular improvement which minimizes the size of the issue as well as the effects into the future.

Appendix Table 1. Margins of Error for Table 4: Effect of 'blanking and allocating' method in 2013 on geographies flagged in 2012 as having a suspiciously high percentage of women aged 15-50 who gave birth in the last year[1]

	Nation	Bartow County, GA	Cherokee County, GA	Floyd County, GA	Durham County, NC	Franklin County, OH	Dallas County, TX	Williamson County, TX
Percent of women who gave birth, 2013	Z	2.2	1.8	2.5	0.9	0.7	0.4	1.0
Percent of women who gave birth, 2012	Z	5.2	3.7	6.2	2.1	0.9	0.6	2.1
Percent of fertility data allocated, 2013	0.1	4.7	3.9	5.9	2.0	0.9	0.6	1.4
Percent of fertility data allocated, 2012	Z	1.4	2.1	2.0	1.4	0.5	0.3	1.3
Percent of women who gave birth--for flagged FR, 2013	X	7.2	3.4	4.3	100.0	1.6	2.3	3.7
Percent of women who gave birth--for flagged FR, 2012	X	10.5	14.7	9.9	9.9	7.9	5.0	11.5
Percent of fertility data allocated--for flagged FR, 2013	X	14.3	9.2	10.7	24.1	7.8	6.2	10.4
Percent of fertility data allocated--for flagged FR, 2012	X	1.6	2.9	81.1	4.7	0.7	1.3	76.9

Z Not zero but rounds to 0.0.

X While overall national level data are included for comparison purposes, flagged FR data was not compiled at the national level since the focus of this table was the effect of individual FRs on specific smaller geographies.

[1] MOE is short for 'margins of error'. Data are based on a sample and are subject to sampling variability. A margin of error is a measure of an estimate's variability. The larger the margin of error in relation to the size of the estimate, the less reliable the estimate. When added to and subtracted from the estimate, the margin of error forms the 90 percent confidence interval.

Source: American Community Survey, 2012 and 2013 1-year data