The Impact of Missing Values on Population Count Estimates in the 2016 National Survey of Children's Health

The 2016 National Survey of Children's Health (NSCH) was conducted in every state and the District of Columbia. All children whose parent or caregiver responded to the survey were included in a final, "raw" data set. From this data set, we can derive counts of individuals who responded to each survey item. These counts are called *"unweighted"* counts since they represent the actual number of surveyed children with a particular response (for example, "yes" or "no").

However, the 2016 NSCH data set also includes *"weights"* for each child. These weights were created by the U.S. Census Bureau, and are used by researchers and the Data Resource Center for Child and Adolescent Health (DRC) to create "population estimates." These population estimates represent not just the children in the survey, but all children in the United States. Both weighted and unweighted data are important, but weighted data are more commonly used, since weights are necessary to represent all children in a given state or the nation.

The purpose of this document is to summarize how the *missing or unknown data* in the raw data set can impact the population count estimates. In previous iterations of the NSCH, there were fewer missing cases or unknown values than those encountered in the 2016 NSCH. The previous versions of the survey were conducted over the phone by an interviewer. The 2016 NSCH was completed by mail or online, with no interviewer involved. This can lead to more items that were left unanswered or skipped. Missed or skipped items can impact the population counts.

Weighted data (or "population estimates")

Weighted data allow researchers and child advocates to use data from the 2016 NSCH to describe not only children whose parent/caregiver responded to the survey, but nearly all children in the United States. By virtue of random sampling, the relatively small number of all children included in the sample dataset can be weighted to represent and estimate the count of nearly all US children. Through weights, the children who are included in the data set become *representative of the population of all non-institutionalized children in the United States who live in housing units (e.g. houses, apartments etc.).*

The population weights in the survey make adjustments to the data for the probability of each child being selected (which varied by state, household, and CSHCN status), survey nonresponse, and to ensure accurate population counts by certain demographic characteristics. For example, the weights included in the 2016 NSCH adjust for the over-representation of certain children (e.g., white) and the under-representation of others (e.g., black) due to different survey response rates. Weights ensure accurate totals for the following child and household-level characteristics: household size, household poverty threshold, educational attainment of the household respondent, race and ethnicity, and special health care needs status by state as well as age and sex nationally. Total child counts by various characteristics are generally estimated using data from the most recent American Community Survey (ACS); in this case, data from 2015. In that survey, there were estimated to be 73,386,395 children 0-17 years old in the United States. The creation of weights in the NSCH 2016 using this information is what allows each child whose parent/caregiver responded to the survey to represent a certain number of children based on actual population distributions for demographic characteristics that they hold.

The impact of missing data on population count estimates

Some types of non-response to the survey are accounted for during the process of creating the weights; for example, the race of non-responders to the survey. When there are non-responses to certain items (e.g., someone skipped the item), then we have *item-level missing data.*

Missing data for certain demographic characteristics, including household poverty which had a high percentage missing (~18%), were imputed as part of the weighting process. In most other cases, items have fewer than 2% missing cases. However, a relatively small missing percentage can still impact the population count estimates even if there is no bias in the population percentage estimate. The tables below present three examples of how different amounts of missing cases can underestimate population counts and offer an adjustment method.

Example 1

Table 1 shows the percentage of children 6-17 years old who were bullied. *Row 1* represents the population count estimate based on the *actual weighted responses received to the survey, removing missing cases from both the numerator and the denominator (i.e., children 6-17 years old with valid bullying responses).* In column D, we see that an estimated 22.6% of children 6-17 years old in the United States were bullied. In column E, (the product of column C multiplied by column D) we see that this corresponds to an estimated 10,981,585 children between the ages of 6 and 17 years old were bullied in the United States (not exact due to rounding error in the population percentage estimate). This is the number that is displayed on the DRC website.

Row 2 represents population count estimates adjusted for missing data by applying the population percentage estimate from known responses to the total population denominator estimate from the survey, regardless of bullying response (i.e., all children 6-17 years). *In row 2, column E, we can see that if there were no missing cases and the population percentage estimate remained the same, then the population estimate would be 11,219,252.*

Population count estimates are still ***estimates***, based upon weights. Practically speaking, in this instance, although the estimates in column E differ between rows 1 and 2 by about 238,000 children, they are both "about 11 million children" in total. In this case, the numbers do not differ greatly once they are rounded (11.0M versus 11.2M).

We recommend reporting weighted counts by rounding up to a single decimal place (e.g. 11.0 million instead of 10,981,585), similar to the level of precision displayed for percentage estimates.

Table 1. Example of the effects of missing values on population count estimates (2.3% missing cases)

	A	B	C	D	E
#	**Item**	**Weighted % missing**	**Population denominator estimate** (Total children 6-17 years old) ✖	**Weighted population %** estimate (**Yes** response) =	**Population numerator estimate** (**Yes** response)
1	Bullied (definitely or somewhat true), age 6-17 years	2.3%	48,518,048 (not including missing cases)	22.6%	10,981,586 (not including missing cases)
2	Bullied (definitely or somewhat true), age 6-17 years	0%	49,642,709 (with zero missing cases)	22.6%	11,219,252

Note: Cells shaded green exclude missing cases. Cells shaded in yellow offer adjusted population count estimates by applying the weighted denominator for all children 6-17 years old.

Example 2

In Table 2, we see the percentage of children 0-17 years old who are reported to live in safe neighborhoods. The number of missing cases is greater in this example (3.0% of the estimated population), which translates to a difference of about 1.4 million children living in safe neighborhoods (45.4M versus 46.8M).

Table 2. Second example of the effects of missing values on population count estimates (3.0% missing)

	A	B	C	D	E
#	**Item**	**Weighted % missing**	**Population denominator estimate** (Total children 0-17 years old)	**Weighted population % estimate** (**Yes** response)	**Population numerator estimate** (**Yes** response)
1	Safe neighborhood (definitely agree)	3.0%	71,162,470 (not including missing cases)	63.8%	45,423,585 (not including missing cases)
2	Safe neighborhood (definitely agree)	0%	73,350,040 (with zero missing cases)	63.8%	46,797,326

Note: Cells shaded green exclude missing cases. Cells shaded in yellow offer adjusted population count estimates by applying the weighted denominator for all children 6-17 years old.

Example 3

Table 3 examines the percentage of children 10-17 years old who are overweight/obese. As compared to the previous example, a larger percent of children (9.0%) did not have valid responses to items needed to ascertain their weight status.

The differences in column E are similar to Example 2 above - that is, the population estimates for overweight/obese differ by about 0.9 million children when accounting for missing cases (9.4M versus 10.3M).

Table 3. Third example on the effects of missing values on population count estimates (9.0% missing)

	A	B	C	D	E
#	**Item**	**Weighted % missing**	**Population denominator estimate** (Total children 10-17 years old)	**Weighted population % estimate** (**Yes** response)	**Population numerator estimate** (**Yes** response)
1	Overweight or obese, age 10-17 years	9.0%	30,059,006 (not including missing cases)	31.2%	9,370,447 (not including missing cases)
2	Overweight or obese, age 10-17 years	0%	33,043,043 (with zero missing cases)	31.2%	10,309,429

Note: Cells shaded green exclude missing cases. Cells shaded in yellow offer adjusted population count estimates by applying the weighted denominator for all children 6-17 years old.

Summary

The number of missing cases for each item are not displayed on the Data Resource Center website; instead, they are removed from analysis and all weighted population percentage and count estimates.

However, it is important to keep in mind:

1) Population count estimates are only that—**estimates.** They are impacted by the number of nonresponses. Population counts without missing data can be obtained from the NSCH or independent sources such as the American Community Survey. Population counts for American Indian/Alaska Native and Native Hawaiian/Other Pacific Islander children should be obtained from independent sources since survey weights did not account for population totals of these groups.
2) We recommend reporting weighted counts by **rounding to nearest hundred thousand** (e.g. 11.0 million instead of 10,981,585) so as not to imply that they are more precise than the prevalence percentage estimates which are also typically presented to a single decimal place.

Citation: Child and Adolescent Health Measurement Initiative (2017). Data Resource Center, supported by Cooperative Agreement U59MC27866 from the U.S. Department of Health and Human Services, Health Resources and Services Administration (HRSA), Maternal and Child Health Bureau (MCHB). "The Impact of Missing Values on Population Count Estimates in the 2016 National Survey of Children's Health". Revised 10/06/17. www.childhealthdata.org www.cahmi.org