The 2012 Annual Homeless Assessment Report (AHAR) to Congress

OCTOBER 2013

Data Collection and Analysis Methodology

The U.S. Department of
Housing and Urban Development
OFFICE OF COMMUNITY PLANNING AND DEVELOPMENT

Table of Contents

A.1   Introduction

This document summarizes the methodology for producing the 2012 Annual Homeless Assessment Report (AHAR). Abt Associates and the University of Pennsylvania (the AHAR research team) developed the methodology.

The 2012 AHAR is based on two primary sources of data:

1. *Homeless Management Information Systems (HMIS).* The HMIS data were collected from a nationally representative sample of communities[1] and cover a one-year reporting period, October 1, 2011 to September 30, 2012. The data contain information on homeless people who used emergency shelters or transitional housing at any point during this period and formerly homeless people who used permanent supportive housing (PSH) programs. HMIS data are unduplicated at the community-level and reported in the aggregate. HMIS data include information on the number, characteristics, and service-use patterns of homeless people. Each AHAR incorporates HMIS data for the most recent, one-year reporting period and compares these data to previous findings. The 2012 AHAR provides comparisons of HMIS data from 2007 to 2012 for all population expect homeless veterans. HUD began collecting HMIS data separately on homeless veterans in 2009.

2. *Continuum of Care (CoC) applications.* The CoC applications provide Point-in-Time (PIT) estimates of homelessness on a single night in January of each year. The PIT data provide estimates of homelessness by sheltered status (sheltered versus unsheltered) and by subpopulation type (chronically homeless people, veterans, and persons with different types of disabling conditions). The PIT data were collected from all CoCs in 2012, and the 2012 AHAR compares these data to previous estimates. The CoC application also provides the number of emergency shelter, safe haven, and transitional housing beds available to serve homeless people, as well as beds in permanent supportive housing programs.

The remainder of this appendix describes the AHAR sample data in more detail. Section A.2 discusses the population represented by the AHAR sample and the information collected about people experiencing homelessness and people using PSH programs. Section A.3 describes how the nationally representative sample was selected and the number of communities that were able to contribute local HMIS data to the AHAR. Section A.4 presents the results of the data cleaning process and describes how useable data were identified for the final AHAR analysis file. Section A.5 describes the process for developing the analysis weights for each site to produce nationally representative estimates.

A.2   Data and AHAR Reporting Categories

This section describes the target population for inclusion in the AHAR sample, the source of data, and the data collection process.

---

[1]   Data from AHAR sample sites is supplemented with data from other Continuums of Care that were not selected as part of the original sample but chose to contribute their HMIS data for the AHAR. These communities are called 'contributing communities'; unlike AHAR sample sites, contributing communities only represent themselves in the national estimates, meaning their data is not weighted to represent other communities to produce the national estimate.

Target Population for the AHAR Sample

The HMIS-based data in the AHAR sample includes information on all people who used an emergency shelter, transitional housing, or permanent supportive housing at any time during a one-year period, from October 1, 2011 through September 30, 2012. The information on emergency shelters and transitional housing programs is then weighted to produce national estimates of sheltered homelessness. The same process is also used to produce national estimates of the number of formerly homeless people who used PSH programs.

The sample does not include individuals who are homeless but live in an area not within a Continuum of Care, or individuals who live in a CoC community but do not use an emergency shelter or transitional housing program. However, given that CoCs cover 97 percent of the U.S. population, including areas with high rates of homelessness, few homeless people are likely to live outside CoC communities. If U.S. Territories are able to provide usable HMIS data they are included in the estimates, however if these territories cannot provide data the research team does not use data from other communities to weight up for them. This year's AHAR estimates include data from Guam and the Virgin Islands but not Puerto Rico. The unsheltered homeless population—people who live on the streets or other places not meant for human habitation—is not represented by the HMIS data in the sample if such people do not use an emergency shelter or transitional housing facility at any time during the one-year data collection period.

One caveat associated with the use of HMIS data for national reporting is that an important subset of homeless service providers is not permitted to participate fully in data collection. The 2005 Violence against Women and Department of Justice Reauthorization Act prohibits "victim service providers"[2] from entering personally identifying information into an HMIS. Even though CoCs were required to include these programs as part of their housing inventory in their funding application, we excluded their beds from our extrapolations; thus, the national estimate of the sheltered homeless population does not include people using residential "victim service" providers.

Homeless Management Information System Data

The information on homeless people in the AHAR sample is based on HMIS data collected by local homeless assistance providers. HMIS are computerized data collection applications operated by CoCs that store data on homeless individuals and families who use homelessness assistance services.

HMIS data have some important features. First, they have been standardized nationally in accordance with HUD's National HMIS Data and Technical Standards Notice (Data Standards).[3] All HUD McKinney-Vento–funded homeless programs are required to collect 14 universal data elements from every client served.[4] The

---

[2]   The term victim service provider is defined as "a nonprofit, nongovernmental organization, including rape crisis centers, battered women's shelters, domestic violence transitional housing programs, and other programs whose primary mission is to provide services to victims of domestic violence, dating violence, sexual assault, or stalking" (72 FR 5056, March 16, 2007).

[3]   69 FR 45888, July 30, 2004.

[4]   Two of the universal data elements (Veterans Status and Disabling Condition) are asked of adults only; two other data elements (Residence Prior to Program Entry and Zip Code of Last Permanent Address) are asked of adults and unaccompanied youth only. Programs that receive Supportive Housing Program (SHP) funding are also required to collect the Program-Specific data elements. Some of these data elements are included in the PSH reporting categories.

Data Standards define each data element. The universal data elements include information on a client's demographic characteristics (e.g., date of birth, ethnicity and race, gender, veteran status, and disability status) and recent residential history (e.g., residence before program entry, program entry and exit dates, and zip code of last permanent address). The data are essential to obtaining an accurate picture of the extent, characteristics, and patterns of service use of the local homeless population.

Second, HMIS data include personally identifying information that allows local communities to produce an accurate unduplicated count of homeless people in their communities. For each person served, programs must collect a client's full name and Social Security Number. The personally identifying information may be used in combination with other client-level information to calculate the number of unique users of homeless services and to identify people who use several types of services.

Third, HMIS data may be manipulated to produce a more comprehensive picture of homelessness when compared to older data collection systems (e.g., paper records). Given that the data are stored electronically in sophisticated software applications, data users may produce cross-tabulations and other outputs that were impractical or impossible before the advent of HMIS. As a result, HMIS data offer new opportunities to study the nature and extent of homelessness.

AHAR Reporting Categories

To facilitate the AHAR reporting process, the AHAR research team developed seven reporting categories that are used to collect information from participating communities. Most of the information required in the reporting categories is based on the universal data elements specified in the HMIS Data Standards.[5] The seven reporting categories are:

1. Individuals served by emergency shelters (ES-IND)
2. Individuals served by transitional housing facilities (TH-IND)
3. Individuals served by permanent supportive housing facilities (PSH-IND)
4. Families served by emergency shelters (ES-FAM)
5. Families served by transitional housing facilities (TH-FAM)
6. Families served by permanent supportive housing facilities (PSH-FAM)
7. A summary table

Reporting categories 1 through 6 contain several sections. The first section is an extrapolation worksheet for estimating the total number of individuals or people in families that used an emergency shelter, transitional housing facility, or PSH program during the data collection period. This section guides the community through a process for estimating the number of individuals or people in families served by providers participating in HMIS as well as by nonparticipating providers. A limited amount of data from the HMIS and communities' bed inventory is required to complete the extrapolation. The remaining sections in each set of reporting categories are designed to capture information about the homeless population served in emergency shelter and supportive housing and the formerly homeless population served in permanent supportive housing in the community. Each set of reporting categories is designed with embedded codes to check for data errors, such as missing values or inconsistent information. A summary sheet of data errors is automatically generated as communities complete the reporting categories, prompting communities to review and correct any errors.

---

[5]   The permanent supportive housing categories collect information on 6 additional data elements.

The summary table captures information on the use of multiple program types during the reporting period. Communities report on the number of people who used both emergency shelter and transitional housing, or were served both as an individual and as part of a family during the reporting period. This information is used to produce the final unduplicated sheltered homeless count, which adjusts for people being counted in multiple program types.

The data submission process is channeled through the Homelessness Data Exchange (HDX), a web-based data collection instrument designed specifically for HUD data collection activities. Communities login to the HDX using a unique username and password and submit the data by either typing the aggregate data into each reporting category or by uploading all their data via an XML schema into the appropriate reporting category. Each community is assigned a data quality reviewer (a member of the research team) who reviews each submission and works collaboratively with representatives from the community to fix any data quality issues. A public version of the HDX is available for viewing and local use: http://sandbox.HUDHDX.info/.

A.3   Sample Selection

This section describes the procedures for selecting a nationally representative sample of 102 jurisdictions for the AHAR.[6]

CDBG Jurisdictions Are Primary Sampling Units

The AHAR uses the geographic areas defined for the allocation of CDBG funds as the primary sampling unit. The four types of CDBG jurisdictions are:

- Principal cities[7]
- Cities with 50,000 or more people (that are not principal cities)
- Urban counties
- Rural areas or non-entitlement jurisdictions

---

[6]   The initial AHAR sample consisted of 80 jurisdictions. Some jurisdictions from the original sample—especially jurisdictions representing rural areas—were unable to provide data to the AHAR because of HMIS implementation issues or other data quality concerns. In addition, several of the rural sample sites did not have any homeless residential service providers located in their jurisdiction. As a result, we were unable to report data by geography. In an effort to improve the scope and quality of data from rural jurisdictions, 22 additional rural jurisdictions were added to the AHAR sample starting with the 2008 AHAR. Thus, there are 102 AHAR sample sites.

[7]   The original file from which the sample was selected used the category of "central city" for CDBG jurisdictions rather than "principal city." However, the CDBG program moved to designation of principal city rather than central city following the OMB guidance, and the definition of central city and principal city are slightly different (see 24 CFR Part 570). Of the 482 CDBG central city jurisdictions that existed both before and after the definition change, 327 central city jurisdictions (68%) became principle cities with the definition change. A small number of non-central cities (85 out of 2,501) in the original file were categorized as principal cities in the 2007 CDBG file. In our analysis by CDBG jurisdiction and in procedures for adjusting the sampling weights, we used the community's current CDBG jurisdiction to ensure that our results accurately represented the current system for designating CDBG jurisdictions.

CDBG jurisdictions constitute the basic building blocks of CoCs. In some cases, the CDBG jurisdiction and the CoC represent the same geographic area (e.g., principal cities are often a single CoC), but, in other situations, the CDBG jurisdiction is a geographic subunit of the CoC (e.g., a small city with 50,000 or more people may be a subunit of a countywide CoC). The selection of 102 CDBG jurisdictions ensures the inclusion of a wide range of sites in the AHAR as well as the reasonably precise measurement of the characteristics of homeless people and their patterns of service use.

The U.S. Department of Housing and Urban Development provided a sampling frame for the selection of CDBG jurisdictions. The sampling frame is a list of all 3,142 CDBG jurisdictions within the 430 CoCs in the 50 states as of 2002. The next section describes the decision to stratify the sites based on geographic type, along with the procedures for selecting certainty and non-certainty sites.

Stratifying the Sample by Type of Geographic Area

A CDBG jurisdiction may be a large principal city of a metropolitan area, a smaller city with a population of 50,000 or more, one or more suburban or urban fringe counties, or a rural area. As such, the number of homeless people in each jurisdiction varies considerably.

Using the relative size of the homeless population in each CDBG jurisdiction to select a sample may increase the precision of the estimates for any particular sample size. However, with the number of homeless people in each CDBG jurisdiction unknown, the study team assumed that the total population in each CDBG jurisdiction provided a measure of relative size of the homeless population for purposes of sample selection. The study team premised the assumption on the likelihood that the number of homeless people is correlated with the total population in the area served by the CDBG jurisdiction. The team further refined the assumption by dividing the sample into strata based on the expected rate of homelessness.[8]

Earlier research on homelessness indicates that the rate of homelessness varies by type of geographic area. For example, Burt (2001) found that 71 percent of the homeless people using homeless-related services are located in principal cities but that only 30 percent of the total U.S. population lives in principal cities.[9] By contrast, rural areas account for 9 percent of the homeless population, but 20 percent of the overall population. Further, suburban/urban fringe areas represent 21 percent of homeless people, but 50 percent of the overall population. These findings suggest that, before using the total population as a proxy for the relative size of the homeless population, the CDBG jurisdictions should be stratified by type of geographic area to account for the fact that the ratio of homeless people to the population varies across geographic areas. Hence, the study team divided the CDBG jurisdictions into four groups based on their classification for the allocation of CDBG funds: principal cities, other cities larger than 50,000, urban

---

[8]   Sampling based on the expected rate of homelessness is an attempt to obtain more precise estimates than those yielded by a simple random sample. If the proxy for the expected rate of homelessness is not correlated with the actual rate of homelessness, the resulting estimates will still be unbiased; however, the extra precision gains go unrealized.

[9]   Burt, Martha. 2001. Homeless Families, Singles, and Others: Findings from the 1996 National Survey of Homeless Assistance Providers and Clients. *Housing Policy Debate,* V12 (4), 737-780. This report presents the share of the homeless population by urban/rural status. The share of the population in each type of geographic area comes from the author's calculations based on March 1996 Current Population Survey data. The results from the Burt study were based on central cities rather than principal cities, but we refer to them as principal cities here because of the high degree of overlap and to make the discussion easier to follow.

counties, and rural areas (i.e., counties that are part of non-entitlement areas). Such stratification increases the precision of estimates.

Very Large CDBG Jurisdictions Selected with Certainty

Given that the size of the population across CDBG jurisdictions is skewed by a few very large jurisdictions covering areas with several million residents, a useful strategy for reducing sampling variability in the estimated number and characteristics of homeless people is to select very large jurisdictions in the sample with certainty. Selecting a CDBG jurisdiction with certainty means that the CDBG jurisdiction represents only itself in the sample estimates but ensures that the sample does not exclude the largest jurisdictions, whose number and characteristics of the homeless population could substantially affect national estimates. Exhibit A-1 lists the 18 CDBG jurisdictions selected with certainty.

For selecting the certainty sites, the study team divided the CDBG jurisdictions into the four geographic-type strata. Assuming the rate of homelessness was the same in each area within each stratum, the study team calculated the standard deviation (square root of the variance) of the number of homeless people for the entire stratum. The team then recalculated the standard deviation by excluding the largest site (as if that site were taken with certainty) to obtain a relative estimate of the reduction in the variance of the estimates that would occur if that site were selected with certainty. In the event of substantial reduction in the variance due to the selection of the certainty unit, the overall variance of the sample estimates will be smaller as the variance contribution to the estimate from the certainty sites is zero. The process of selecting the next-largest site as a certainty site continued until the reduction of the variance or standard deviation was small or marginal. The process resulted in the identification of 11 certainty sites consisting of eight principal cities, one other city larger than 50,000, and two urban counties (but no non-entitlement areas).

Based on earlier research findings showing that homeless people are disproportionately located in principal cities, the study team identified 7 additional principal cities as certainty sites, for a total of 15 principal cities in the certainty sample (and 18 certainty sites in total). The team selected the seven additional principal cities with certainty because the cities had among the largest populations of people living in emergency and transitional shelters in the 1990 and 2000 Census counts.[10] All seven certainty sites had one of the 10 largest counts in either 1990 or 2000.[11] Given that so many homeless people live in these cities, it is important to include them with certainty in a nationally representative sample.

---

[10]   For 1990 counts, see U.S. Department of Housing and Urban Development. "Allocating Homeless Assistance by Formula." A Report to Congress, 1992. For 2000 counts, see U.S. Census Bureau. "Emergency and Transitional Shelter Population: 2000." A Census 2000 Special Report.

[11]   The other 8 certainty sites in principal cities were all ranked in the top 15 in the 1990 or 2000 Census counts.

Exhibit A-1: Geographic Characteristics and Population of 18 Certainty Sites

	Geographic Area	Type of CDBG Entity	Size of Housed Population	Census Region	CoC Name
1	NEW YORK CITY	Principal City	8,008,278	Northeast	New York City Coalition/CoC
2	LOS ANGELES	Principal City	3,694,820	West	County of Los Angeles, CA
3	CHICAGO	Principal City	2,896,016	Midwest	Chicago CoC
4	HOUSTON	Principal City	1,953,631	South	Houston/Harris County
5	PHILADELPHIA	Principal City	1,517,550	Northeast	City of Philadelphia
6	PHOENIX	Principal City	1,321,045	West	Maricopa CoC
7	SAN DIEGO	Principal City	1,223,400	West	City of San Diego Consortium
8	DALLAS	Principal City	1,188,580	South	Dallas Homeless CoC
9	DETROIT	Principal City	951,270	Midwest	City of Detroit CoC
10	SAN FRANCISCO	Principal City	776733	West	City and County of San Francisco
11	BOSTON	Principal City	589,141	Northeast	City of Boston
12	WASHINGTON, DC	Principal City	572,059	South	District of Columbia Homeless Services
13	SEATTLE	Principal City	563,374	West	Seattle-King County CoC
14	CLEVELAND	Principal City	478,403	Midwest	Cuyahoga County/Cleveland CoC
15	ATLANTA	Principal City	416,474	South	Atlanta Tri- Jurisdictional
16	LOS ANGELES COUNTY	Urban County	2,205,851	West	County of Los Angeles, CA
17	COOK COUNTY	Urban County	1,712,784	Midwest	Cook County CoC
18	ISLIP TOWN	City >50,000	322,612	Northeast	Suffolk County CoC Group

Note: CDBG jurisdiction type and the population of each jurisdiction are as of 2002 when these sites were identified as certainty sites for the sample and were taken from a file HUD provided called "COC_GeoAreasInfo.xls".

Selection of Non-Certainty Sample

There are currently 102 AHAR sample sites. The selection of the non-certainty sites occurred in two phases. Phase one was completed in 2005 and included 62 non-certainty sites. The 62 non-certainty sites and the 18 certainty sites (80 total sample sites) constituted the original sample for the 2005, 2006, and 2007 AHARs. Phase 2 was completed for the 2008 AHAR and added 22 non-certainty sites to the original sample.

Phase 1: Selecting 62 Non-Certainty Sites. To select the 62 non-certainty sites for the original sample, the study team divided the 3,124 CDBG jurisdictions into 16 strata based on the four types of geographic

areas and Census regions. As discussed earlier, the team divided the sample into strata based on the type of geographic area because earlier research indicated that the rate of homelessness is higher in principal cities than in other areas. The team further divided the sample into Census regions because business cycles might affect regions differently and result in variation in rates of and trends in homelessness across regions. Dividing the sample into strata that are more similar in terms of the rate of homelessness and the characteristics of homeless people than the overall population reduces the variance of the sample estimates for a particular sample size. Stratified sampling also eliminates the possibility of some undesirable samples. For example, with a simple random sample, one possible sample might include sites only in rural areas or sites only in the Northeast, both of which are undesirable samples.

One possibility considered for the non-certainty sample was allocation of the sample to the stratum in proportion to the population in each stratum. However, such an approach ignores the research indicating that a disproportionate share of the homeless is located in principal cites. Ignoring information on the location of the homeless population would lead to a relatively high degree of imprecision in national estimates such that 20 of the 62 non-certainty sites would be allocated to principal cities, 6 to non–principal cities, 16 to urban counties, and 20 to rural areas. The same number of rural areas as principal cities would be selected even though earlier research suggests that only 9 percent of the homeless population lives in rural areas whereas 70 percent lives in principal cities.

Another possibility under consideration for the non-certainty sample was allocation of the total non-certainty sample of 62 CDBG jurisdictions to each of the 16 strata in proportion to the adjusted population in each stratum, where the adjustment accounts for different rates of homelessness across geographic areas. This allocation method produces the highest degree of precision of national estimates for a given sample size. The adjusted population is the population of people living in an area multiplied by an adjustment factor for the expected rate of homelessness in that area. With the rate of homelessness in principal cities roughly five times that of other areas,[12] the study team multiplied the population in principal cities by five so that the adjusted populations would reflect the relative number of homeless people expected in each stratum. If the adjusted population were used to allocate the non-certainty sites across the strata, 39 of the 62 original non-certainty sample sites would have been allocated to principal cities, 4 to non– principal cities, 8 to urban counties, and 11 to rural areas. While optimal for national estimates, the number of sites in the non–principal city stratum was too small for subnational estimates.

The sampling allocation procedure ultimately used for AHAR data collection strikes a balance between the most precise national estimates possible with a sample of 62 non-certainty sites and reasonably sized samples from each of the four types of geographic areas. The study team allocated the 62 original non-certainty sample sites across the 16 strata based on the square root of the adjusted population. The result is a sample allocation between the allocation in proportion to the population and the allocation in proportion to the adjusted population. Accordingly, 27 of the 62 original non-certainty sites are in principal cities, 8 are in non-principal cities, 13 are in urban counties, and 14 are in rural areas. The allocation means lower variances of the estimates than either simple random sampling or sampling in

---

[12]   The ratio was determined as follows. Burt (2001) found that 71 percent of the homeless population lived in central cities in 1996. At the same time, Current Population Survey data indicate that only 30 percent of the overall population lived in central cities at that time. The ratio of the share of the homeless population to the share of the overall population in central cities is 2.36. The ratio is 0.42 for non– principal city portions of Metropolitan Statistical Areas and 0.46 for rural areas. Dividing the principal city ratio by the rural ratio (2.36/0.46) equal 5.1, suggesting that the rate of homelessness is about five times higher in central cities than in rural areas.

direct proportion to the population and provides better representation of non– principal city areas than the allocation in proportion to the adjusted population.

To select the non-certainty sites in each stratum, the study team divided the sites into groups based on size and then randomly selected one site from each group. The number of non-certainty sites allocated to each stratum determined the number of groups, and each group in a stratum contained the same number of sites. Sampling from groups based on population size is beneficial in that it ensures that the sample has a similar distribution of CDBG jurisdiction sizes as the population. Given that the size of the homeless population is expected to correlate with the total population within strata, similarity in distribution is an important feature of the sample.

Phase 2: Adding 22 Rural Non-Certainty Sites. The data collection results from the 2005-2007 AHAR reports indicated that many rural communities (or non-entitlement CDBG areas) did not have emergency shelters or transitional housing programs located in these jurisdictions. Among the few rural sample sites that did have emergency shelters and/or transitional housing programs, many of those programs were not entering data into an HMIS. As a result, previous AHAR reports did not capture information from many rural jurisdictions, and the lack of data increased the variance of the AHAR estimates and made the analysis of rural/suburban versus urban homelessness less reliable.

In 2008, 22 new rural communities were added to the AHAR sample, increasing the total number of rural jurisdictions to 36 and the total number of AHAR sample sites to 102. The 22 AHAR sample sites that were added in 2008 were selected in the same manner as the original non-certainty sample sites. The original 2002 sampling frame of 3,142 CDBG jurisdictions within the 430 CoCs in the 50 states was used to select the new rural communities. However, the original file was compared with an updated 2006 CDBG list of jurisdictions to remove from the sampling frame jurisdictions that had either merged with other jurisdictions since 2002 or had changed their status from non-entitlement (rural) areas to entitlement areas.

The sample was stratified to ensure that each of the four census regions was represented. The goal was to select at least three rural communities from each census region that had at least one emergency shelter or transitional housing program. In some cases, more than three communities for a particular region were selected if inventory information reported by CoC suggested that the communities did not have any emergency shelters or transitional housing programs. That is, from each region, we randomly selected rural jurisdictions until we had at least three rural jurisdictions with at least one emergency shelter or transitional housing program. In total, 22 new rural sample sites were added in 2008; three from the Northeast region; seven from the South region; seven from the Midwest region; and five from the West region.

The final AHAR sample contains 102 sample sites, and Exhibit A-2 shows the total number of certainty and non-certainty sites selected from each region-CDBG type stratum. The sample sites contain over 40 million people, or approximately 16 percent of the population living within CoC communities and 14 percent of the U.S. population. The expectation is that the sample will contain an even higher proportion of the U.S. homeless population because the selection procedures intentionally oversampled areas with a high rate of homelessness (i.e., principal cities). About two-fifths of the selected sites (42 sites) are principal cities, even though only one-third of the total population lives there. The other 60 sample sites were distributed across the three remaining CDBG jurisdictions: non-principal cities with a population over 50,000 (9 sites), urban counties (15 sites), and non-entitlement/rural areas (36 sites).

Exhibit A-2: Number of Sites in Universe and Sample by Region-CDBG Type				
Stratum	Number of Geographic Areas in Universe	Number of Certainty Sites in Sample	Number of Noncertainty Sites in Sample	Total Sample
Northeast Principal City	86	3	5	8
South Principal City	151	4	8	12
Midwest Principal City	124	3	7	10
West Principal City	106	5	7	12
Northeast City >50,000	81	1	2	3
South City >50,000	48	0	2	2
Midwest City >50,000	55	0	1	1
West City >50,000	114	0	3	3
Northeast Urban County	33	0	3	3
South Urban County	54	0	4	4
Midwest Urban County	33	1	3	4
West Urban County	34	1	3	4
Northeast Non-entitlement County	148	0	6	6
South Non-entitlement County	812	0	11	11
Midwest Non-entitlement County	890	0	11	11
West Non-entitlement County	373	0	8	8
Total	**3,142**	**18**	**84**	**102**

Note: For sampling, HUD provided a file called "COC_GeoAreasInfo.xls" with a list of 3,219 CDBG jurisdictions, jurisdiction type, and population of each jurisdiction. Geographic areas in the U.S Territories and Puerto Rico and three duplicate records were eliminated, resulting in a sampling frame of 3,142 CDBG jurisdictions. In addition, four CDBG areas in Massachusetts and one in New Hampshire included overlapping geographic areas and double-counted the population; therefore, the population was evenly divided across the overlapping CDBG jurisdictions before sampling.

Addition of Contributing Sites

In addition to the 102 sample sites selected for the study, many other communities nationwide volunteered to provide data for the report to help produce more precise national estimates. The additional communities are entire Continuums of Care and are termed "contributing sites." In the 2012 AHAR, 342 contributing communities provided data for use in the AHAR report. As with the sites selected with certainty, data from the contributing sites represent themselves in the national estimates.

A.4   AHAR Data Cleaning

This section presents the data cleaning results for the AHAR. For each AHAR sample site and contributing community, the study team reviewed each reporting category (e.g., ES-IND) for reporting irregularities, focusing on three indicators:

- HMIS-bed coverage rate
- Average daily bed utilization rate
- Proportion of missing variables

Bed Coverage Rate

HMIS-bed coverage rate refers to the proportion of beds in a community that participate in HMIS. The HMIS-bed coverage rate is equal to the total number of HMIS-participating beds divided by the total number of beds in a community. The indicator is important because the accuracy of the extrapolation technique depends on obtaining reasonably high bed coverage rates.[13] The study team evaluated each reporting category on its own merits—that is, calculated an HMIS-bed coverage rate for all six reporting categories separately—and excluded from the final AHAR analysis any reporting category with an HMIS-bed coverage rate below 50 percent.

Average Daily Bed Utilization Rate

Average daily bed utilization rate refers to the frequency of bed use on an average day. It is equal to the number of homeless people who use a program on an average day during a specified period divided by the total number of year-round equivalent beds[14] in the current inventory during the same period. Utilization rates above 100 percent typically indicated missing exit dates in the HMIS; unusually low utilization rates often suggested that providers did not enter data on all clients served into HMIS. In situations where unusually high or low utilization rates could not be explained or confirmed as accurate by the community, the study team excluded from analysis all data from the reporting category.

Proportion of Missing Variables

Missing data limit the ability to present a complete picture of homelessness. Exhibit A-3 presents the proportion of missing values for the weighted 2012 emergency shelter and transitional housing AHAR data. The data element most constrained by missing values was length of stay in prior living arrangement, which was missing for 12.5 percent of adult clients. Though still a high rate, 2012's rate of missing is considerably lower than in the 2010 AHAR (20.6 percent) and the 2011 AHAR (14.8 percent). Missing rates for all of the demographic data elements were below 3 percent. The missing data rates for disability status (3.8 percent), veteran status (1.5 percent), prior living situation (3.9 percent), and ZIP code of last permanent address (5.7 percent) continued the decline seen in earlier AHARs.

Exhibit A-4 shows the proportion of missing values among the weighted 2012 permanent supportive housing AHAR data. Among the data elements that applied to all three program types, the permanent supportive housing missing data rates were similar to the emergency shelter and transitional housing missing rates. 2012's permanent supportive housing rate of missing was consistently lower than in the 2011 AHAR with ZIP code of last permanent address showing the largest percentage point drop (from 13

---

[13]   Before releasing the AHAR reporting requirements, the study team tested the extrapolation procedures with data from Philadelphia and Massachusetts under a variety of coverage rate assumptions, taking a random sample of providers (to match 50, 75, and 90 percent HMIS bed-coverage rates) and comparing the extrapolated estimates to the true population counts for these jurisdictions. The findings show that extrapolation estimates were accurate for HMIS bed-coverage rates above 50 percent and were more precise with higher coverage rates. The threshold of an HMIS bed-coverage rate of 50 percent was as representative as possible of a set of participating sample sites. (See 2004 National HMIS Conference Breakout Session Materials "Extrapolation Methods" for more information on the extrapolation testing, available at www.onecpd.info.)

[14]   A year-round equivalent bed counts seasonal beds as partial beds in direct proportion to the length of the covered period for which the provider makes the bed available. For example, a bed from a provider with a seasonal bed open in January, February, and March would count as one-fourth of a bed since the reporting period is 12 months.

to 7 percent). The permanent supportive housing data collection included 6 additional data elements that were not collected for emergency shelter and transitional housing (elements 17 to 22). Among these, one data element, the destination at program exit, had a high proportion of missing data (11.4 percent).

Exhibit A-3: Proportion of Missing Values In Emergency Shelter and Transitional Housing (weighted data), 2012			
Variable	Percent Missing	Variable	Percent Missing
1. Gender of adults	0.1	9. Household type	0.2
2. Gender of children	0.2	10. Living arrangement before program entry	3.9
3. Ethnicity	1.5	11. Length of stay in earlier living arrangement	12.5
4. Race	2.4	12. ZIP code of last permanent address	5.7
5. Age	0.3	13. Number of nights in program (adult females)	0.1
6. Household size	0.0	14. Number of nights in program (adult males)	0.1
7. Veteran status	1.5	15. Number of nights in program (female children)	0.1
8. Disability status	3.8	16. Number of nights in program (male children)	0.1

Exhibit A-4: Proportion of Missing Values In Permanent Supportive Housing (weighted data), 2012			
Variable	Percent Missing	Variable	Percent Missing
1. Gender of adults	0.1	12. ZIP code of last permanent address	6.7
2. Gender of children	0.1	13. Number of nights in program (adult females)	0.2
3. Ethnicity	0.7	14. Number of nights in program (adult males)	0.1
4. Race	1.9	15. Number of nights in program (female children)	0.2
5. Age	0.4	16. Number of nights in program (male children)	0.3
6. Household size	0.1	17. Type of disability	3.7
7. Veteran status	2.9	18. Length of most recent consecutive stay (adult female)	0.1
8. Disability status	3.3	19. Length of most recent consecutive stay (adult male)	0.2
9. Household type	0.2	20. Length of most recent consecutive stay (female children)	0.2
10. Living arrangement before program entry	4.5	21. Length of most recent consecutive stay (male children)	0.2
11. Length of stay in earlier living arrangement	11.3	22. Destination at program exit	11.4

The study team did not exclude reporting categories from the AHAR analysis file because of missing data. Instead, the estimates are based on non-missing data, and the team has marked estimates in the AHAR report based on data elements with missing rates over 20 percent.

Based on the data-quality indicators, the study team classified all sample sites and the contributing communities into five categories describing the usability of their AHAR data. Exhibit A-5 summarizes the findings. Overall, 436 communities participated in the AHAR, including 94 sample sites and 342 contributing communities. 190 communities (41 sample sites and 149 contributing communities) provided usable data across all six reporting categories; 224 communities (31 sample sites and 193 contributing communities) submitted usable data for only some of their reporting categories; and 22 sample sites had no emergency shelter, transitional housing, or permanent supportive housing providers.[15]

Exhibit A-5: 2012 AHAR Participation Status of Sample and Contributing Communities				
Status	**Total**		**Number of Sample Sites**	**Number of Contributing Communities**
	Percentage	**Number**		
Participating in the AHAR				
All table shells	39	190	41	149
Partial table shells	46	224	31	193
Complete Zero Providers	5	22	22	-
Subtotal	90	436	94	342
Not Participating in the AHAR				
Submitted unusable data	4	19	2	17
No data submitted	6	28	6	22
Subtotal	10	47	8	39
Total Communities	100	483	102	381

In total, 8 of the 102 sample sites (8 percent) were unable to participate in the AHAR, in most cases because implementation issues prevented the site from producing information from their HMIS. A few of the sites were far enough along to submit data but were still working through implementation problems or had recently made major changes to their system that raised questions about the data quality. The study team judged data to be unusable if the bed coverage rate was below 50 percent; if the bed utilization rates were unreasonably high/low and could not be properly explained; if the community contact expressed concern over data accuracy; or if the other quality control procedures raised issues that site staff could not rectify.

The 2012 AHAR witnessed a year-over-year increase of 11 communities contributing useable data (from 425 in 2011 to 436 in 2012).[16] Moreover, the number of usable reporting categories (among emergency shelter and transitional housing categories) increased from 1,108 in the 2011 AHAR to 1,168 in the 2012 AHAR. (Exhibit A-6 shows the number of usable reporting categories for the 2012 AHAR.) In total, there were 1,053,187 person-records reported across the AHAR reporting categories (882,199 across emergency shelter and transitional housing and 170,988 in permanent supportive housing) that were used to generate the national estimates.

---

[15] These sites still contribute to the national count of homelessness because they represent other communities with no providers.

[16] The total number of communities in the country dropped from 484 in the 2011 AHAR to 483 in the 2012 AHAR due to CoC mergers.

Exhibit A-6: Number of Usable Reporting Categories by Program-Household Type, 2012			
Program-Household Type	**Total**	**Sample Sites**	**Contributing Communities**
Emergency shelters for individuals	269	40	229
Transitional housing for individuals	281	49	232
Emergency shelters for families	292	48	244
Transitional housing for families	326	52	274
Subtotal for ES and TH	*1,168*	*189*	*979*
Permanent supportive housing for individuals	331	52	279
Permanent supportive housing for families	300	49	251
Total	*1,799*	*290*	*1,509*

Note: The tallies include only the reporting categories where the site has providers in a given category and provides usable data. The table does not include the zero provider categories.

Between 2007 and 2012, there was a large decrease in the proportion of homeless people that were missing information on where they lived prior to entering a shelter. The improvement in data quality ironically led to misleading percentage changes when comparing this information across reporting years. That is, many more homeless people appeared to be coming from the various prior living situations than in the past, producing large percentage changes even though homelessness declined during this period. The large changes were being produced by the decline in missing rates—i.e., unknown living situations became known. The following steps were taken to address this reporting issue:

1. The percentage distribution among known people was applied to the count of unknown people. For example, if 200 people were missing prior living status and 10 percent of known people were living in a place not meant for human habitation, 20 of the unknown people would be added to the count of people living in a place not meant for human habitation.

2. Because the breakdown of prior living arrangement is unique to each household and program type, percentage distributions were applied separately to each household and program type.

A.5   AHAR Weighting and Analysis Procedures

This section describes the process of obtaining national estimates from the raw HMIS data submitted by participating communities. The estimates of the number and characteristics of the homeless population using homelessness services are based on weighted data. The study team designed the sampling weights to produce nationally representative estimates from the sites that provided data. The steps for obtaining the final estimate are listed here and described in more detail below.

- **Step 1**:   Staff from the AHAR sample and contributing sites filled out reporting categories with information (raw data) from emergency shelters and transitional housing providers that had entered data into their local HMIS.

- **Step 2**: The raw data were adjusted by reporting category within each site to account for providers that did not participate in the site's HMIS.

- **Step 3**: Base sampling weights were developed for all selected sites based on the assumption that 100 percent of the AHAR sample sites provided information.

- **Step 4**: Base sampling weights were adjusted to account for contributing sites.

- **Step 5**: Weights were adjusted for nonresponse to determine the preliminary analysis weights.

- **Step 6**: Weights were further adjusted to correct for stratum with zero usable sample beds and to reduce large outlier weights.

- **Step 7**: A final adjustment factor was derived to account for people who used more than one type of homeless service provider.

- **Step 8**: National estimates were calculated by using the final weight (Step 6) and the final adjustment factor (Step 7).

People using PSH programs are no longer homeless because they are living in permanent housing. Therefore, these data were not included in the sheltered homeless estimates. However, the same weighting process was used to produce separate national estimates of the number and characteristics of people using PSH programs during the reporting period.

Step 1: Staff from AHAR sites filled out reporting categories with information from emergency shelter, transitional housing, and permanent supportive housing providers that had entered data into their local HMIS.

Communities participating in the AHAR logged into the HDX and entered the information (raw data) on the number of homeless people, their characteristics, and their patterns of service use. The information was reported separately for each reporting category: individuals using emergency shelters (ES-IND); people in families using emergency shelters (ES-FAM); individuals using transitional housing (TH-IND); people in families using transitional housing (TH-FAM); individuals using permanent supporting housing (PSH-IND); and people in families using permanent supportive housing (PSH-FAM). The information was then aggregated into a seventh set of tables, the summary tables, to provide total cross-program estimates for the site. The aggregated set of summary tables is downwardly adjusted for any homeless people staying in more than one program-household type during the study period (i.e. people are unduplicated).

Step 2: The raw data were adjusted by reporting category within each site to account for providers that did not participate in the site's HMIS.

Where participation in the HMIS was less than 100 percent, the raw data at each site were upwardly adjusted to account for nonparticipating providers (i.e., providers that did not submit data to HMIS). This adjustment, or extrapolation, was carried out separately by reporting category within each site. The extrapolation technique assumes that nonparticipating providers serve the same number of unique people per available bed as participating providers during the study period. It makes a small adjustment for the overlap between users of participating and nonparticipating providers.[17]

---

[17] Given that data from nonparticipating providers were not available, it is impossible to verify this assumption. However, it is the most reasonable assumption in that it is accurate when nonparticipating providers are missing

The post-extrapolation results for each site are estimates of the total number of people served by each reporting category across the entire site, including non-participating providers, during the study period.

Step 3: Base sampling weights were developed on the assumption that 100 percent of the AHAR sample sites provided information.

The study team selected the largest sites (i.e., the CDBG jurisdictions with the largest populations) with certainty. As such, each site's base sampling weight is 1.0, meaning that each respective site's data represent only that site. The study team divided the noncertainty sites into 16 strata based on the four Census regions (East, West, Midwest, and South) and four CDBG types (three types of entitlement communities—principal city, urban county, other city with population greater than 50,000—and one type of non-entitlement community). The base sampling weights for the noncertainty sites are the number of shelter beds available in each stratum divided by the number of shelter beds in sample AHAR communities in each stratum. For example, if there were 100 beds located in sites in a stratum and 10 beds were in sites selected as part of the sample, the base sampling weight for selected sites in that stratum would be 10. Each noncertainty site in a stratum had the same chance of being selected as part of the sample; therefore, each site within a stratum has the same weight.

If all the sample sites provided full AHAR data (in the absence of contributing sites), national estimates of the homeless population would be calculated by multiplying each site's base sampling weight by the extrapolated number of people with each characteristic at the site and then aggregating across sites.

Step 4: Base sample weights were adjusted to account for contributing sites.

Three hundred and forty two communities volunteered to provide their HMIS-based data for the 2012 AHAR. The data from these contributing communities increase the reliability of the AHAR estimates. The study team treated all of these sites as certainty sites and assigned them a weight of 1.0 such that each site would represent only itself in the national estimates. The study team adjusted the base sampling weights of the noncertainty sites downward to represent only the noncontributing sites in their respective stratum. For example, assume that the sample sites in a stratum included a total of 10 beds and that the base weight was 10 (there are 100 beds in the stratum: 10 sample beds times a weight of 10 equals 100). If the contributing sites included 10 beds in that stratum, the sample weight would be downwardly adjusted to 9. In other words, the sample sites originally represented 100 beds in their stratum, but, with the contributing sites now representing 10 of those 100 beds, the sample sites need to represent only 90 beds. The addition of the contributing sites did not affect the base sampling weight (1.0) of the certainty sites.

If all the sample sites and contributing sites provided full AHAR data, the study team would calculate national estimates of the homeless population by multiplying each site's base weight by the extrapolated number of people for each characteristic (e.g. gender, age, race, etc.) collected at the site and then aggregate across sites.

---

at random or at least not systematically missing in a way correlated with the number of people they serve per available bed.

Step 5:   The base weights were adjusted for nonresponse to derive the preliminary analysis weights.

The above base weights assume that all the sample and contributing sites provided data for all reporting categories except for those for which they have no providers in their jurisdiction. However, 8 sample sites were not able to provide any usable data, and 31 other sample sites were unable to provide data for all their reporting categories (i.e., they provided partial data). One hundred ninety three contributing sites also provided only partial data. In addition, 22 sample sites had no providers (i.e., no emergency shelter, transitional housing, or permanent supportive housing programs). The 'zero provider sites' are part of the estimate (because they represent themselves and all nonsample zero provider sites in the population) but need to be treated differently from the other sites. Once the study team confirmed that the site had no providers, it needed no further information. Given that the zero provider sites did not have any information for the AHAR reporting categories, all were considered respondents.

Recognizing that some participating sites provided only partial data (i.e., data on some but not all of their reporting categories) that was useful for the AHAR report, the study team carried out the nonresponse adjustment to the weights separately for each of the six reporting categories. That is, each site contributing data to the AHAR has six analytic weights—one for each reporting category. However, for any reporting category for which a site was not able to provide data, the analytic weight is zero. The respondent sites for that reporting category represent the site. (Step 8 describes the procedure for aggregating across reporting categories to arrive at national estimates.)

Below is a description of how the weight for each type of site was adjusted for nonresponse to derive the final analysis weights.

(a) The weights of the ***contributing sites*** did not change; each contributing site continued to represent itself with an analytic weight of 1.0 for each program-household type for which it provided data.

(b) The weights of the ***zero provider sites*** did not change. Their weight remained the base weight calculated in Step 4 because all zero provider sites in the sample are considered respondents. In essence, the zero provider sites produced a response of 100 percent. Stated differently, since none of the *non-response* sites has zero providers, the zero provider sites would not appropriately represent them.

(c) For the ***certainty sites*** providing data, base weights were adjusted so that the analytic weights represented all certainty sites. The adjustment was made separately for each program-household type within four weighting classes based on region: North, South, East, and Midwest. [18] The nonresponse adjustment was based on the relative number of shelter beds in the nonrespondent sites and accounts for the possibility of a high degree of size variation among certainty sites. The nonresponse adjustment formula follows:

---

[18]   Fifteen of the 18 certainty sites are principal cities; therefore, the nonresponse adjustment essentially occurs within CDBG type.

Total number of beds within a reporting category at certainty sites in region	÷	Number of beds within reporting category at respondent certainty sites in region

For example, assume that six of the seven certainty sites in the West provided TH-IND data and that one site did not. If the nonrespondent certainty site had 1,000 TH-IND beds and the six participating certainty sites had 5,000 beds, the weight of the six participating certainty sites would be multiplied by 6/5 (6,000 divided by 5,000). The adjustment assumes that the nonrespondent certainty sites would serve approximately the same number of people per bed as the participating certainty sites. The nonresponse adjustment for certainty sites was derived separately by region based on the judgment that homeless providers in principal cities in the same region were more likely than principal cities overall to serve people with similar characteristics.

(d) For the *noncertainty sites*, the weights of the participating sites were upwardly adjusted to represent all the sites meant to be represented by the nonrespondent sample sites. The adjustment was carried out separately for each program-household type within 16 weighting classes based on type of region and CDBG jurisdiction: (1) principal city, (2) city with greater than 50,000 population, (3) urban counties, and (4) and nonentitlement areas. The nonresponse adjustment was the same as that used for certainty sites--the ratio of total number of beds in sample sites within the weighting class divided by number of beds in participating sample sites. The adjustment was then multiplied by the base weight to create the final weight.

There was a net increase in AHAR participation among sites from the 2011 AHAR to the 2012 AHAR. This rise was particularly large among contributing sites that fully participated in every reporting category (133 contributing sites participated fully in the 2011 AHAR compared to 149 in the 2012 AHAR). The number of sample sites and contributing sites that were not able to provide any useable data fell in the 2012 AHAR (from 59 total sites in the 2011 AHAR to 47 total sites in the 2012 AHAR). Increases in participation led to a smaller nonresponse adjustment in 2012, thus lowering the final weight.

Step 6:   Weights were further adjusted to correct for stratum with zero usable sample beds and to reduce large outlier weights.

The AHAR sample was divided into 16 strata based on census region and CDBG type. Wherever possible, the research team used data from sample communities within a stratum to weight up for communities that did not provide usable HMIS data. However, in some cases the research team received insufficient data from sample sites within a particular stratum. In these cases, the stratum was combined with the most similar available stratum within the same program type, so that the usable sample beds within the most similar available stratum would represent nonreporting beds from both strata. For example, among emergency shelters for people in families, none of the AHAR sample communities selected in the northeastern cities greater than 50,000 stratum had participating emergency shelter programs for families. However, there are 829 noncertainty, noncontributing beds within that stratum that needed to be represented. To account for these beds, the 829 nonreporting beds were added to the Midwest/cities with greater than 50,000 people stratum, and the reporting sample sites located in the Midwest/cities with greater than 50,000 people represented all nonreporting, noncertainty beds in nonentitlement areas located in the Northeast and the Midwest.

After correcting for stratum with zero reporting beds, there were some large weights that caused sites to contribute a disproportionate number of people to the final estimate. This occurred when there were only a small number of reporting sample beds, but a large number of nonreporting beds within the stratum. In these cases, a slight difference in the sample site from the nonreporting sites would cause a significant bias in the weighted national estimates. To address this problem, outlier weights were combined with the most similar available stratum (which did not have an outlier weight) within the program/household category in order to decrease the effect of the outlier weight. For instance, there were 1,892 total noncertainty TH-IND beds in nonentitlement areas in the Midwest, but there were only 4 reporting sample beds, yielding a non-response adjusted weight of 473. To reduce this weight, the Midwest/non-entitlement area stratum was combined with the Midwest/urban county stratum so that all reporting sample beds in the Midwest non-entitlement areas and urban counties represented all nonreporting sample beds in the Midwest non-entitlement areas and urban counties.

Step 7:   Final adjustment factor was derived to account for users of several program and household types.

To calculate national estimates that require data aggregation across the four reporting categories, an adjustment is needed for people who used more than one program-household type during the study period. People can enter emergency shelter or transitional housing within the reporting year and can be in shelter as an individual or in a family as household compositions alter. For example, if a person used an emergency shelter for individuals and then a transitional housing program for individuals during the reporting year, the person will appear in more than one set of reporting categories for the study period; aggregation of the numbers from the four emergency shelter or transitional housing or among the two permanent supportive housing reporting categories would double count that person.[19] The needed adjustment is the same type of adjustment embedded in the AHAR summary table for sites providing data on all four emergency shelter transitional housing reporting categories. For the 239 participating sites (69 sample sites and 170 contributing communities) providing data on all four emergency shelter and transitional housing reporting categories, the adjustment factor was the actual adjustment factor calculated from how much overlap the sites reported with their HMIS data. However, for the 193 participating sites that provided only partial data, it was not possible to calculate the overlap adjustment factor from their data. Instead, for all partial reporting sites, the study team used the average overlap adjustment factor from the 239 sites providing full data. Thus, for partial reporting sites, the overlap adjustment factor was assumed to be 0.9503.

Separately, using the same methods, the study team calculated an overlap factor for people who used both permanent supportive housing for individuals and permanent supportive housing for families. Where a community only contributed one of the two permanent supportive housing categories, the average overlap factor among sites contributing data for both permanent supportive housing household types (0.9994) was used.

---

[19]   The adjustment was done separately for emergency shelter/transitional housing and permanent supportive housing, since people served in permanent supportive housing are not considered homeless. (Permanent supportive housing programs are for "formerly homeless" people.) Multi-program type estimates of homelessness only include emergency shelter and transitional housing.

The overlap adjustment factor was calculated as follows:

Total unduplicated number of people served at the full-reporting sites	÷	Total number of people served at the full-reporting sites before accounting for people served by more than one program-household type

Step 8: Calculate national estimates.

To calculate national estimates, the study team first calculated the total number of people with each characteristic within each site for each the six reporting categories. Then, within each reporting category, the team multiplied the final analysis weight (from Step 7) for each site by the number of people for each reporting characteristic (e.g., gender, age, race, etc.) in that site's reporting category. Next, the team summed the number of people in each site across sites to arrive at the estimated number of people with that characteristic who were served in that reporting category. For estimates of the number of people served by all four emergency shelter and transitional housing reporting categories or the two permanent supportive housing categories, the team summed totals across the four reporting categories and then multiplied by the adjustment factor from Step 7. Percentage calculations followed the same procedures by calculating both the numerator and denominator of the desired percentage.