U.S. Department of Health & Human Services

2017 Annual Survey of Refugees Data File User's Guide

A Technical Research Manual

Prepared by Timothy Triplett
Carolyn Vilter
Urban Institute

July 2019

Table of Contents

Introduction

Since the 1980s, the Office of Refugee Resettlement[1] (ORR) has conducted the Annual Survey of Refugees (ASR), which collects information on refugees during their first five years after arrival in the U.S. The ASR is the only scientifically-collected source of national data on refugees' progress toward self-sufficiency and integration. ORR uses the ASR results alongside other information sources to fulfill its Congressionally-mandated reporting requirement following the Refugee Act of 1980.

In the spring of 2018, ORR completed its 51st Annual Survey of Refugees (ASR). The data from the ASR offer a window into respondents' first five years in the United States and show the progress that refugee families made towards learning English, participating in the workforce, and establishing permanent residence. This user's guide presents basic information on the 2017 ASR public use data file.

The first section of the user's guide gives an overview of the survey, including descriptions of the sample design and data collection procedures. The next section discusses the structure of the ASR data file and describes the variables included on the data file. Section 3 explains how missing data were coded, and it includes recommendations on how to handle the missing data when conducting analysis. Sections 4 and 5 provide information on how to use the survey weights and procedures for calculating variances and standard errors of survey estimates. Section 6 provides recommendations for users who plan to compare 2017 estimates to earlier ASR estimates or plan to combine the 2017 and 2016 ASR public use data files. The appendices to this guide include data dictionaries that display both weighted and unweighted frequencies and a copy of the English version of the questionnaire.

[1] The Office of Refugee Resettlement (ORR) at the Administration for Children and Families in the U.S. Department of Health and Human Services (HHS) serves refugees and other humanitarian entrants, including asylees, Cuban and Haitian entrants, Special Immigrant Visa holders, Amerasians, victims of human trafficking, and unaccompanied children. By providing these arrived populations with critical resources, ORR promotes their economic and social well-being. Of these populations, the Annual Survey of Refugees focuses solely on refugees who have come to the U.S. in the past five fiscal years.

The goal of this guide is to provide users with sufficient technical information about the data and the survey to properly access and analyze the public use data. Some subsections are preceded by an icon that informs the reader about the nature of the subsequent material. Three are used, and they appear below along with an explanation:

 *Indicating critical points that all users should understand*

 *Indicating useful tips, but not essential reading*

 *Indicating sections meant primarily for advanced users*

Section 1: Overview of the Design of the Survey

The 2017 ASR design replicated the 2016 ASR design[2], which used a full cross-sectional national sample of refugees entering within the past five years. This section documents the research design, data collection and data processing protocols. It also presents outcomes (e.g., sample sizes) and paradata results such as response rates.

The population of interest – the study population – for the 2017 ASR is defined as refugees entering the U.S. between FY 2012 and FY 2016, inclusive, who are at ages 16 and over at the time of the 2017 ASR interview[3]. Because the interviews were conducted in early 2018, the population includes a small number of refugees younger than 16 at the time of arrival to the U.S.

While this covers five distinct fiscal years of refugee entrants, there is special policy/analytic interest in collapsing years into three domains as follows:

- Cohort 1 – Refugees entering FY 2012 and FY 2013,
- Cohort 2 – Refugees entering FY 2014 and FY 2015, and
- Cohort 3 – Refugees entering FY 2016

Table 1 shows the distribution of the study population by fiscal year as well as cohort. About 353,000 refugees (of all ages) entered the U.S. in FY 2012-2016, with roughly equal numbers arriving annually between FY 2013 and FY 2015. FY 2016 had about

[2] In ASR surveys prior to 2016, the ASR survey design was longitudinal, consisting of a cross-sectional sample of refugees arriving one year prior to the study and surveyed that year and followed for four subsequent waves, totaling five annual surveys.

[3] Note that the ASR data files include person records of children under 16 at the time of interview, non-refugees, and refugees who entered outside of the FY eligibility window. However, only a small number of demographics (e.g., age, sex) were collected for these cases. The full set of substantive measures (e.g. language proficiency, education, labor force participation, etc.) were collected for eligible refugees age 16 or over at the time of interview.

Table 1: Population Distribution of Refugees Arriving Between 2012 and 2016

	Fiscal Year of Arrival	Number of Refugees*	% of Refugees
Cohort 3	2016	84,994	24%
Cohort 2	2015	69,933	20%
	2014	69,987	20%
Cohort 1	2013	69,926	20%
	2012	58,238	16%
Total		**353,078**	**100%**

* Source: FY 2012-2016 data compiled from Department of State admissions reports

85,000 refugee arrivals, while a smaller number entered four years earlier in FY 2012 (about 58,000). These refugees represent 130 countries and 208 non-English languages.

"Refugees" are persons, not households. However, when refugees come to the U.S., they often enter with their family members. For an entering refugee family, there is a Principal Applicant (PA) whose refugee case is the basis for admission. This person is often the head of the household. Table 2 shows the distribution of PAs entering the U.S. between FY 2012-2016 by family size at arrival. Just under half of the roughly 148,000 PAs had families of two or more people.

Table 2: Principal Applicants — Cohorts 1-3		
Family Size	***%***	***Cum %***
1	51%	51%
2	12%	63%
3	12%	76%
4	11%	86%
5	6%	92%
6	4%	96%
7+	4%	100%
No. of Principal Applicants	**147,891**	

The 2017 ASR targeted 1,500 completed interviews from refugee *households* entering the U.S. between FY 2012-2016. The sample was designed to allow for separate estimates and analyses from each of the three designated cohorts. Moreover, the design needed to accommodate both household- and person-level analyses.

The sample was drawn as fresh cross sections by cohort; there was no longitudinal component. The survey objectives required that – in addition to primary stratification by cohort – the sample of households (i.e., PAs) be stratified at least by year of entry and geographic region of origin.

The 2017 ASR sampling frame was ORR's Refugee Arrivals Data System (RADS) dataset.

Sample Design

The 2017 ASR employed a stratified probability sample design of refugees. The first stage of selection was the household (PA) and the second stage was the selection of persons within households. Principal features of the sample design are highlighted below.

Sample Allocation to Cohorts.
The ASR design targeted equal numbers of household interviews by cohort. This is depicted in Table 3, which shows an allocation of 500 households per cohort. This means that there was an oversample of households for FY 2016, the most recent year of entry. This allocation prioritizes the statistical precision to cohorts.

Table 3: 2017 ASR Target Number of Household Interviews by Cohort and Year of Arrival

	A	B	C	D
	Cohort Household Population %	2017 Target HH Interviews by Cohort	Expected Interviews by Arrival Fiscal Year	Arrival Fiscal Year
Cohort 3: FY 2016	22%	**500**	500	2016
Cohort 2: FY 2014-2015	43%	**500**	250	2015
			250	2014
Cohort 1: FY 2012-2013	35%	**500**	275	2013
			225	2012
Total	100%	1,500	1,500	

Respondent Selection.
The ASR can be used for both household-level and person-level analyses. Although the Principal Applicant represented the household sampling unit, data were collected by proxy from all eligible refugees aged 16+ within each sampled household. Eligible refugees are those PA household residents who entered between FY 2012-2016 and are 16 years old or older on the day of the

2017 ASR interview. Thus, the 2017 ASR sample design featured household-level element sampling and person-level clustered sampling. The PA served as the proxy informant for all eligible refugees within the household.

Population Coverage and Language Diversity.
An important design issue involved addressing the 200+ languages associated with the ASR population. Analysis of RADS data suggests that only 5 percent of cohort 2-3 refugees speak "good" English, suggesting that the clear majority of interviews needed to be in languages other than English. Table 4 tabulates primary language spoken by refugees using RADs data. We see that just over 70 percent of refugees speak 12 non-English languages, while about 75 percent of refugees speak one of 17 non-English languages. Unfortunately, it takes 238 languages to fully cover all refugees.

The 2017 ASR was offered in 16 non-English languages (17 including English) identified in Table 4 corresponding to rows 1 to 17. This achieved an overall coverage of 75 percent of the FY 2012-2016 refugee population.

Table 4: Coverage of 2017 ASR Refugees' Primary Languages

Language Count	Primary Spoken Language	Primary Spoken Language Cum %	Primary Spoken Language %
1	**Arabic**	23%	23%
2	**Nepali**	35%	13%
3	**Somali**	46%	10%
4	**Sgaw Karen**	51%	6%
5	**Spanish**	55%	4%
6	**Kiswahili**	59%	4%
7	**Kinyarwanda**	62%	2%
8	**Farsi, Western**	64%	2%
9	**Burmese**	66%	2%
10	**Tedim**	67%	2%
11	**Chaldean***	69%	2%
12	**Tigrinya**	71%	2%
13	**Lai**	72%	1%
14	**Russian**	73%	1%
15	**English**	74%	0.4%
16	**French**	74%	0.4%
17	**Amharic**	75%	0.4%
18-238	Remaining 221 languages	*Not Covered for 2017 ASR*	

**** Interviewing in Chaldean will only be available via an interpreter.***

Stratification.
Within each of the three cohort strata, the following factors were used for stratification: Cohort, year of arrival (for Cohorts 1 and 2 only), geographic region, native language, age group, gender, and family size at arrival (1, 2, 3+ persons). Proportionate stratified samples were drawn independently within cohort.

Accounting for Nonresponse.
Past ASR studies have been subject to highly differential survey nonresponse rates due to the compounding effects of following participants from the previous year's study (in addition to newly entering cohorts), the difficulty of tracing, and the inability to conduct surveys in every language. To address this important design issue, we modeled our proposed sample by cohort using 2016 ASR tracing efficacy and nonparticipation rates.

Table 5 shows the sampling strategy to reach an expected 1,500 completed interviews from an original sample of roughly 6,000 refugees, of which just under 2,000 would be located. On average 6,015/1,500 = 4.0 sampled PA refugees would be needed to produce a completed interview, assuming the 2017 ASR field experience is like that of the 2016 ASR.
The sampling strategy was to sample 6,000 using the proportional allocations to Cohorts shown in Table 5. Moreover, an additional sample of 4,500 was drawn but held in reserve. The reserve sample was available to be released as needed to supplement the original sample to attain the targeted 1,500 completed interviews.

Table 5: Expected Sample Sizes and Dispositions for the Proposed 2017 ASR

Cohort	Fiscal Years of Arrival	Expected Tracing Effectiveness	Sample Drawn	Successfully Traced & Contacted	Respondent Participation	Expected Interviews
Cohort 3	**2016**	40%	1622	654	76%	500
Cohort 2	**2014-15**	33%	2008	665	75%	500
Cohort 1	**2012-13**	28%	2385	668	75%	500
Total			**6,015**	**1,987**		**1,500**

 ***Replicated Samples***
In the 2017 ASR, replicated samples were used to create the replicate weights.

Design Summary.
Principal features of the final sample design are summarized in Table 6.

Table 6: Summary of 2017 ASR Sample Design Elements	
Design Issue	**Design Approach**
Survey Population Definition	Refugees aged 16 years or older at the time of interview who arrived in the U.S. during FY 2012-2016
Cohort Definition	Cohort 1: FY 2012-2013 arrivals Cohort 2: FY 2014-2015 arrivals Cohort 3: FY 2016 arrivals
Sampling Frame	RADS dataset
Sampling Unit	Refugee Households, achieved by sampling Principal Applicants (PAs)
Sample Allocation to Cohorts	Equal allocation of 500 households to each cohort
Population Coverage	Refugees in the ASR from only the languages covered by the translations plus Chaldean (interpreter only), yielding a 75% refugee population coverage
Stratification	Cohort, year of arrival, geographic region, native language, age at arrival, gender, and family size at arrival (1, 2, 3+ persons)
Accounting for Nonresponse	Expect to use 6,000 households to produce 1,500 completed household interviews
Sample Release	Field a sample of 6,000; hold a reserve sample of an additional 4,500 in case some is needed to achieve the targeted 1,500 completed interviews
Respondent Selection & Interviewing	Use household selection to collect data on the PA, the PA's household, and all eligible adults aged 16+ within a household via proxy reporting by the PA

Survey Administration

The survey administration procedures for the 2017 ASR are detailed in this section.

 ***Sample Management.***
The 2017 ASR employed a sample management plan integrating the sample design and field protocols to include locating subjects, contacting them and conducting telephone interviews. A sample of 6,006 PAs was released at the start of data collection. A reserve sample of about 4,500 was held in case some portion was needed to meet the interview target of 1,500.

Translation of Materials.
For the 2017 ASR, revisions to the 2016 survey instruments and materials were translated into 16 different languages, including English. The survey retained an interpreter to conduct interviews in a 17th language, Chaldean. The only changes to the questionnaire in 2017 involved streamlining the household roster and demographic questions to improve efficiency. Instead of structuring the household roster (qn1a) around the head of household, it was structured around the respondent. Subsequent demographic questions were asked of respondents first, and repeated for other household members only if their responses were reported to be different (qn1gaa, qn1haa, qn1iaa). These changes are highlighted in the questionnaire document in Appendix A.

The languages that were translated and available in CATI (Computer-assisted telephone interviewing) and hard copy (written only) form appear in Table 7 below. As described above, these languages cover about 75 percent of the eligible adult refugee population.

Table 7: 2017 ASR Languages Available by Translation Mode						
Language Count	Primary Spoken Language	Translation Mode	ASR Refugee Cum %	ASR Refugee* %	% Normed to 17 Selected Languages	Cum % Normed
1	Arabic	CATI	19%	19%	25%	25%
2	Nepali	CATI	36%	17%	22%	47%
3	Somali	CATI	46%	10%	13%	60%
4	Sgaw Karen	CATI	53%	8%	10%	70%
5	Spanish	CATI	58%	5%	7%	77%
6	Burmese	CATI	60%	2%	3%	79%
7	Farsi, Western	CATI	63%	2%	3%	82%
8	Kiswahili	CATI	65%	2%	3%	85%
9	Tedim	Written	67%	2%	3%	87%
10	Tigrinya	Written	69%	2%	3%	90%
11	Lai	Written	71%	2%	3%	92%
12	Kinyarwanda	Written	73%	2%	3%	95%
13	Chaldean	Interpreter	75%	2%	3%	98%
14	***English***	***CATI***	***75%***	***0.4%***	**1%**	**98%**
15	***Russian***	***Written***	***75%***	***0.9%***	**1%**	**99%**
16	***Amharic***	***Written***	***76%***	***0.5%***	**1%**	**100%**
17	***French***	***Written***	***77%***	***0.3%***	**0%**	**100%**
* Source: RADS data						

Field Protocols.

In this section, we detail the protocols involved in fielding the Annual Survey of Refugees, beginning with managing the sample using paradata.

 ***Managing the Sample.***

Weekly progress reports were prepared and reported such statistics as:

- Percentage of sample by the amount and type (if any) of updated information obtained;
- Percentage of sample released, pending, and finalized;
- Percentage of sample by all intermediate and final dispositions;
- The sample's net yield (i.e., average number of sampled units per completed interview);

- Number of calls made, refusals incurred, and interviews completed;
- Demographics of completed interviews vs. entire sample; comparisons by respondent demographics (language, sex, country of origin, family size);
- Completed interviews by source of contact information.

 ***Tracing Sampled Subjects.***
The RADS data included contact information for most of the sample. The vast majority of the sample had contact information that was 2 to 6 years old. The entire sample underwent tracing to secure as much updated location information as possible.
Our tracing protocol was implemented on the entire sample (i.e., both the initial release and the reserve sample) at the beginning of fielding for the sake of efficiency. This included the use of National Change of Address as well as TransUnion batch updates.

 ***Letters of Introduction.***
Upon the release of the sample into the field, an introduction letter containing a $2 advance incentive was issued via first class postal service. Seventeen versions of this letter were prepared, and the letter sent to the sampled refugee was tailored to their specific primary language spoken as reported in RADS. A research study logo was used to visually "brand" the survey and make it easier for refugees to distinguish ASR letters from junk mail or bills. The introductory letters themselves appeared in two languages — English plus the primary language spoken by the PA — and contained a call-in number that would allow the respondent to communicate in their primary spoken language as well as offering call-in options for their likely second and third languages, when applicable. It also contained an ASR-specific email address so that the refugee could communicate questions and/or updated contact information. The letter also contained a postage-free, mail-back form for updating the refugee's contact telephone number.

 ***Outreach to Resettlement Agencies.***
Outreach to resettlement service providers was made via email contact. The communications informed State Refugee and Health Coordinators of the fielding of the survey and requested that they share the information with community-based service providers likely to interact with refugees presenting questions about the letter of introduction or phone call inquiries.

 ***CATI Programming & Testing.***
The hard copy questionnaire was programmed and tested to ensure proper flow and appropriate skip logic. The CATI program included nine distinct languages as shown under the "Translation Mode Column" of Table 7 (see languages corresponding to rows containing the word "CATI").

 ***Hiring and Training of Interviewers.***
ASR interviewers underwent a four-hour study-specific training in addition to the typical generic training undertaken by all interviewers. The study-specific training protocol covered orientation on refugee issues and the U.S. refugee resettlement system. It also covered securing survey participation, asking sensitive questions and averting refusals, topics of cultural sensitivity, refusal aversion techniques, and the intricacies of the survey questionnaire itself. The training included participation in multiple mock interviews in English and non-English languages.

 ***Quality Control.***
Quality control is an important part of ensuring data quality. About 7 percent of interviewer hours were 'live-monitored' to ensure fidelity to the protocol. As needed, interviewers who failed to follow procedures were re-trained or released, depending on the nature of the departure from protocol.

 ***Post-Participation Fulfilment Protocols.***
The 2017 ASR provided post-participation incentives (a $25 gift card) via first class mail.

Specification of Field Period.
Tracing commenced in December 2017, letters of introduction were issued early January, and calling began the second week in January. The survey data collection period lasted 12 weeks, from January 10 to April 7, 2018.

Conducting Interviews.
The CATI sample management system executed a calling protocol that required ten call attempts per sampled subject across different times of day and different days of the week. For a given sampled subject, calling was deliberately spread over a couple of weeks including a rest for about one week before resuming dialing (as needed after the first five attempts and provided the short field period allows for a full week of "rest"). Also, whenever an updated telephone number was obtained, the calling algorithm was reset to allow a fresh set of ten call attempts.

At two points in the field period, decisions were made regarding whether to release additional sample to achieve the 1,500 completed interview target. The decision points were:

(a) at the commencement of data collection;

(b) at 4 weeks into data collection; and

(c) at 8 weeks into data collection.

A high degree of attention to cultural sensitivity and relevance was integrated into our field protocols. This included matching interviewer and subject gender to prevent male interviewers from calling female subjects. Moreover, religious holidays and other important calendar dates were loaded into the CATI sample management system so that interviewers did not call refugees on solemn religious holidays.

Results

The 2017 ASR field effort resulted in 1,515 completed refugee household/PA interviews. Table 8 presents the final dispositions from our sample of 6,006 Primary Applicants at the end of the field period. Final completed household interviews from the three cohorts (i.e., FY 2016, FY 2014-15, FY 2012-13) came within 12 percent of the desired targets of 500 per cohort.

Response Rates.

An overall response rate of 25 percent was achieved. The response rate was driven by the ability to locate and speak to (1515+534)/6006 = 32 percent of the sample, meaning that two thirds of the sample could neither be located nor (if located) successfully contacted.

The overall response rates decreased with time since arrival to the U.S., varying from 18 percent for FY 2012-13 refugees to 26 percent for FY 2014-15 refugees and a high of 34 percent for FY 2016 refugees.

The second set of rows in Table 8 ("Screened Refugee, Not Interviewed") presents detailed dispositions among those who were contacted and verified, yet did not participate in the survey. Just over a quarter of contacted refugees, 537/(1515+537) = 26 percent, were contacted but did not participate; this type of noncooperation varied little by cohort.

The bottom set of rows of Table 8 (Unable to Find) shows difficulty in locating refugees.

Table 8: 2017 Annual Survey of Refugee Final Dispositions

2017 ASR Final Dispositions	FY 2012-2013		FY 2014-2015		FY 2016		TOTAL	
Disposition:	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Total Sample	***2,380***	***100%***	***2,006***	***100%***	***1,620***	***100%***	***6,006***	***100%***
Completed Interview	**440**	**18%**	**527**	**26%**	**548**	**34%**	**1,515**	**25%**
Screened Refugee, not interviewed	**206**	**9%**	**176**	**9%**	**155**	**10%**	**537**	**9%**
Refusal after screener	*36*	*17%*	*33*	*19%*	*22*	*14%*	*91*	*17%*
Breakoff	*84*	*41%*	*67*	*38%*	*53*	*34%*	*204*	*38%*
Callbacks (Screener Completed)	*52*	*25%*	*49*	*28%*	*51*	*33%*	*152*	*28%*
Answering machine	*21*	*10%*	*17*	*10%*	*18*	*12%*	*56*	*10%*
Physically or mentally unable/incompetent	*1*	*0%*	*1*	*1%*	*4*	*3%*	*6*	*1%*
Do not call (Final Refusal)	*12*	*6%*	*9*	*5%*	*7*	*5%*	*28*	*5%*
Unable to Screen Refugee (Located)	**897**	**38%**	**647**	**32%**	**406**	**25%**	**1950**	**32%**
Always busy	*32*	*4%*	*23*	*4%*	*15*	*4%*	*70*	*4%*
No answer	*428*	*48%*	*301*	*47%*	*164*	*40%*	*893*	*46%*
Answering machine-don't know if household	*223*	*25%*	*127*	*20%*	*94*	*23%*	*444*	*23%*
Call blocking	*26*	*3%*	*34*	*5%*	*24*	*6%*	*84*	*4%*
Housing unit, unknown if eligible respondent	*110*	*12%*	*78*	*12%*	*55*	*14%*	*243*	*12%*
Callbacks (No Screener Completed)	*71*	*8%*	*70*	*11%*	*38*	*9%*	*179*	*9%*
No screener completed Other	*7*	*1%*	*14*	*2%*	*16*	*4%*	*37*	*2%*
Unable to Find Refugee (Not Located)	**837**	**35%**	**656**	**33%**	**511**	**32%**	**2004**	**33%**
Fax/data line	*1*	*0%*	*0*	*0%*	*1*	*0%*	*2*	*0%*
Non-working number	*441*	*53%*	*291*	*44%*	*191*	*37%*	*923*	*46%*
Business, government office, other	*4*	*0%*	*4*	*1%*	*5*	*1%*	*13*	*1%*
No eligible respondent	*47*	*6%*	*47*	*7%*	*46*	*9%*	*140*	*7%*
Sample without address and phone number	*4*	*0%*	*3*	*0%*	*0*	*0%*	*7*	*0%*
Insufficient contact information	*340*	*41%*	*311*	*47%*	*268*	*52%*	*919*	*46%*

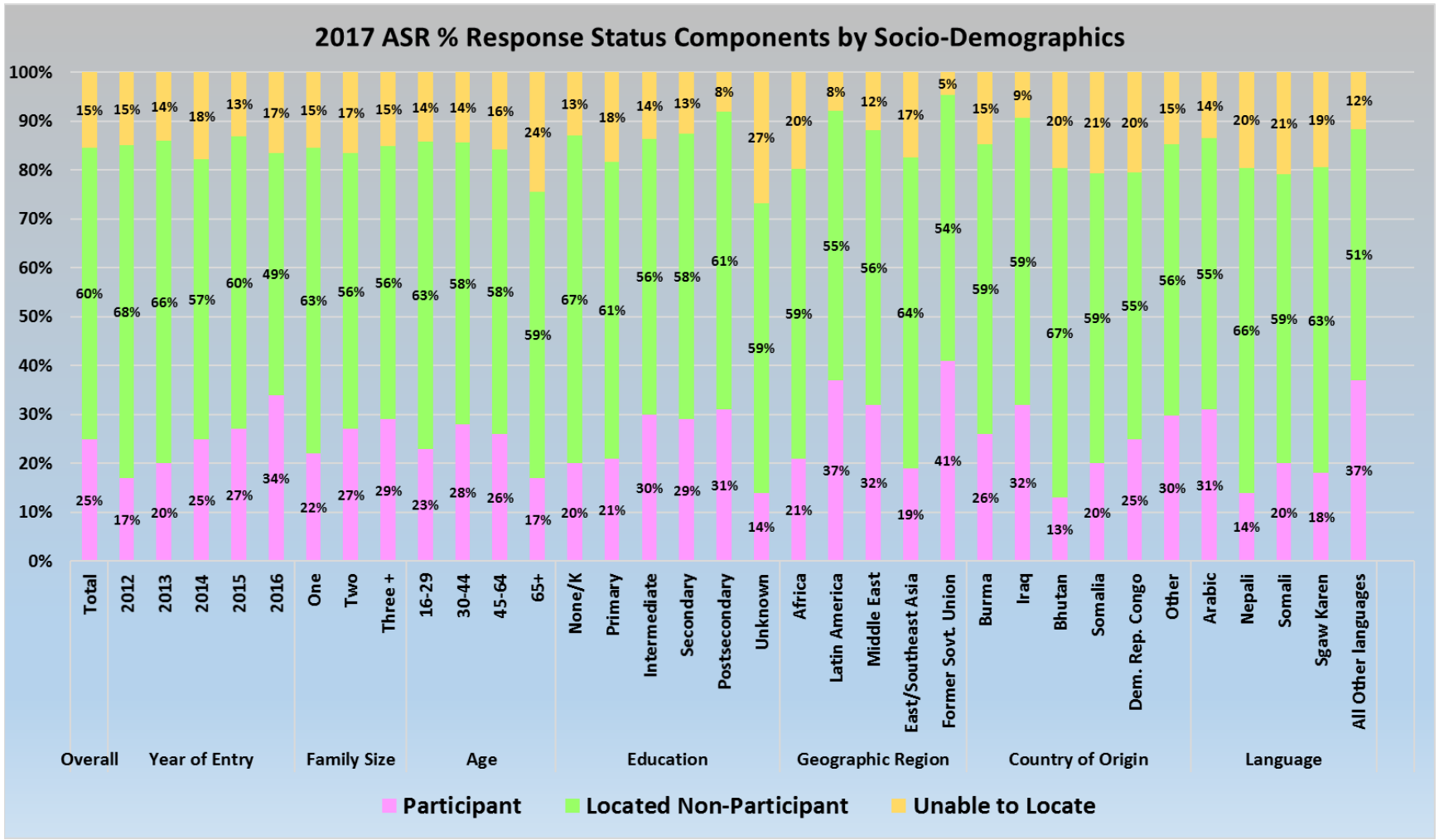
 ***Nonresponse Analysis.***
The variation in response rate components across selected demographic variables appears in Figure 1. The bottom of the graph shows the overall response rate in pink across a variety of demographic factors. Rates of 'unable to contact' appear in yellow at the top of the graph for these subgroups, and percentages of the sample 'located non-participant' in the middle bars (green). Note that unable to contact means that there ultimately was no number to dial to verify the eligibility and conduct an interview of the sampled PA. It includes 7 cases for which there was no contact information in the RADS (so no tracing could be done) and another 919 cases with insufficient information to call (i.e., the bottom two rows of Table 8). The label 'located non-participant' denotes all remaining PA nonparticipants and represents those for whom some calling was attempted.

Overall response rates. The leftmost bar shows an overall response rate of 25 percent. Overall response rates across subgroups shown visually reveal the monotonic increase in response rate by recency of arrival. A similar but not as striking monotonic trend appears for family size – the larger the household, the more likely it was to locate and interview the sampled refugee. The graph also suggests that response rates were highest among middle aged subjects 30 to 64 years old; lowest response rates occurred for young adults and the elderly. Response rates were also associated with levels of education at the time of entry. Generally, the higher the refugee education level at arrival, the higher the participation rate. Considerable variation in response rates was seen in terms of the refugee geographic region, country of origin, and language, which are correlated characteristics. Refugees from Bhutan and Somalia experienced the lowest response rates, 13 and 20 percent, respectively. The highest response rate was seen for Iraq (32 percent).

Turning to the top portion of the graph, showing the percentage of the sample 'unable to locate,' there is not much variation by recency of arrival (ranging from 13 to 18 percent). There is little variation by family size, as well. Some demographic subgroups that were particularly difficult to locate included elderly refugees 65 years old or over (24% unable to locate), refugees with no education (27% unable to locate), and refugees from Bhutan, Somalia, and the Democratic Republic of Congo (20-21% unable to locate).

 Section 4 of this manual explains how to properly weight the data to get correct person-level or household-level estimates and provides a few examples.

Figure 1: ASR Response Rate Analysis



Section 2: Types and Definitions of Variables on the Data File

The ASR data are organized into a person-level file where each person has one record Household-level and administrative variables have been attached to each person's data record. This data structure was chosen because it is consistent with the way the data are collected in the survey and accommodates both person and household-level analysis.

It is important to understand that there are 3 types of person records included in the data file: 1) Persons who are not refugees who came to the U.S. during the past five years. These people are included on the data file because they live in the household, but they do not have person-level weights because they are not "eligible refugees" and are generally not included in any of the analyses; 2) Refugee children who are 15 years of age or younger at the time of survey administration. These individuals have person-level weights, but very little data was collected on them, so they are also usually not included in the analyses; 3) Refugees who are 16 or older at the time of survey administration who came to the U.S. during the past five years. These individuals have person-level weights and a full set of person measures that are either self-reported (in the case of respondents) or proxy reported (in the case of respondents' household members).

The ASR data file contains four types of variables:

1. **Survey variables** store information obtained directly from questions asked on the survey. The variable name for each survey variable begins with the letters "qn" and corresponds with the question number from the survey questionnaire. The questionnaire can be found in Appendix A.

2. **Constructed variables** summarize or combine information from survey variables. We have included in this dataset and user's guide only constructed variables that aggregate information from several survey variables to create more complex measures. Data users should check how constructed variables can meet their analytic needs before going directly to the use of survey variables, especially if they believe that the measure of interest involves multiple survey items. The variable name for each constructed variable begins with the letters "ui." The constructed variables in the data file are all described in this section of the user's guide.

3. **Administrative variables** provide information that was not obtained directly from a respondent, such as their geographic location, or information

about the interviewing process itself, such as language of the interview. Administrative variables include identifier variables, such as person or household ID. The administrative variables in the data file are all described in this section of the user's guide.

4. The variable name for each **weight variable** begins with the word "weight." For more information on weights, see chapter 4 of this guide.

 *The ASR has a complex survey design. To produce unbiased estimates from the 2017 ASR data, it is critical that researchers use the survey weights.*

Constructed Variables

ui_lfp: This variable reports individuals' labor force participation status: in the labor force, not in the labor force, or doesn't know or refused to respond. It was created using responses to qn5a and qn13. Individuals are considered "in the labor force" if they report working at a job anytime the week before survey administration (qn5a) or looking for work during the four weeks before survey administration (qn13). Individuals are considered "not in the labor force" if they report not working at a job anytime the week before survey administration (qn5a) and not looking for work during the four weeks before survey administration (qn13) (or answer "don't know" or refuse to respond to qn13). Respondents who either don't know or refuse to respond to both qn5a and qn13 are marked "Don't know and/or refused" for *ui_lfp*.

ui_emprate: This variable reports individuals' employment status: employed, unemployed, not in the labor force, or doesn't know or refused to respond. It was created using responses to qn5a and qn13. Individuals are considered "employed" if they report working at a job anytime the week before survey administration (qn5a), "unemployed" if they report not working at a job anytime the week before survey administration (qn5a) and looking for work during the four weeks before survey administration (qn13), and "not in the labor force" if they report not working at a job anytime the week before survey administration (qn5a) and either report not looking for work during the four weeks before survey administration, don't know, or refuse to respond (qn13). Respondents who either don't know or refuse to respond to qn5a are marked "Don't know and/or refused" for *ui_emprate*.

ui_medicaidrma: This variable reports individuals' receipt of Refugee Medical Assistance (RMA) or Medicaid: receives RMA/Medicaid, doesn't receive RMA/Medicaid, or doesn't know or refused to respond. It was created using responses to qn29c and qn29d. Individuals are designated "Receives RMA/Medicaid" if they select "Medicaid or Refugee Medical Assistance" in response to qn29d. Individuals are designated "Does not receive RMA/Medicaid" if they select any qn29d response option(s) excluding "Medicaid or Refugee Medical Assistance," or if they answer "Not covered in any month" in response to qn29c. Respondents who either don't know or refuse to respond to both qn29d are marked "Don't know and/or refused" for *ui_medicaidrma*.

ui_lpr: This variable reports individuals' legal permanent residency (LPR) status and plans: has already adjusted LPR status, has not applied to adjust LPR status but plans to, has not applied to adjust LPR status and does not plan to, or doesn't know or refused to respond. It was created using responses to qn27a and qn27c. Individuals are designated "Already adjusted LPR status" if they report having applied to adjust their immigration status to LPR (qn27a) and designated "Plans to adjust LPR status in future" if they report not having applied to adjust their status

(qn27a) but planning to in future (qn27c). Individuals are designated "Not applied to adjust, may not" if they report not having applied (qn27a) and not planning to (qn27c); report not having applied (qn27a) and answer "don't know" or refuse to answer (qn27c); or answer "don't know" to qn27a but select a response option for qn27c. Respondents who either don't know or refuse to respond to both qn27a and qn27c are marked "Don't know and/or refused" for *ui_lpr*.

ui_school: This variable reports individuals' educational pursuits in the United States: pursuit of a high school degree, associate's degree, bachelor's degree, master's/doctorate, professional school degree, certificate/license, other form of education, or doesn't know or refused to respond. It was created using responses to qn25a and qn25c. The variable reports responses to qn25c, with the additional step of flagging as "None" individuals who report not attending school in the United States (qn25a) and flagging as "Don't know and/or refused" individuals who answer "don't know" or refuse to answer qn25a or qn25c. Note that "certificate/license" was not a specific option in the questionnaire; the "certificate/license" counts come from coding related "other, specify" responses into a new response option.

ui_agect_arrival: This is a categorical variable that reports individuals' grouped ages at arrival in the United States. It was created using responses to qn1d and qn1jyear. Given that the survey was administered in 2018, the year respondents reported an individual arriving in the U.S. (qn1jyear) was subtracted from 2018 to find years in the U.S. This value was subtracted from individuals' reported ages (qn1d) to find their age at arrival in the U.S. Finally, this value was grouped into categories: less than zero (0), zero to seventeen (1), eighteen to twenty-four (2), twenty-five to thirty-nine (3), forty to fifty-four (4), and fifty-five and up (5). Respondents who either don't know or refuse to respond to qn1d are marked "Don't know and/or refused" for *ui_agect_arrival*.

ui_cashassist: This variable reports households' receipt of cash assistance: receives cash assistance, doesn't receive cash assistance, or doesn't know or refused to respond. It was created using responses to qn31a, qn32a, qn33a, and qn34a. A respondent's household is designated "Receives cash assistance" if they report one or more persons in their household receiving TANF (qn31a), Refugee Cash Assistance (RCA) (qn32a), Supplemental Security Income (SSI) (q33a), or General Assistance (GA) (qn34a) in the twelve months before survey administration. Households whose respondent either doesn't know or refuses to respond to all four questions (qn31a, qn32a, qn33a, and qn34a) are marked "Don't know and/or refused" for *ui_cashassist*. Remaining households are designated "Does not receive cash assistance."

ui_soi_pubassist: This variable reports households' receipt of public assistance: receives public assistance, doesn't receive public assistance, or doesn't know or refused to respond. It was created using responses to qn30a, qn31a, qn32a, qn33a, qn34a, and qn38a. A respondent's household is designated "Receives public

assistance" if they report one or more persons in their household receiving food stamps (qn30a), TANF (qn31a), Refugee Cash Assistance (RCA) (qn32a), Supplemental Security Income (SSI) (q33a), or General Assistance (GA) (qn34a) or residing in public housing (qn38c) in the twelve months before survey administration. Otherwise, if more than two responses to the public assistance questions were missing, households were marked "Don't know and/or refused" for *ui_soi_pubassist*. Households that reported not receiving any of the public assistance programs and had two or fewer missing responses were designated "Doesn't receive public assistance."

ui_soi: This variable reports households' source(s) of income: receives earnings, receives public assistance, receives both, does not receive either, receives public assistance but missing earnings data, receives earnings but missing public assistance data, doesn't receive public assistance but missing earnings data, or doesn't know or refused to respond. The variable was created using responses to qn18c(a-e), qn30a, qn31a, qn32a, qn33a, qn34a, and qn38a. A respondent's household is designated "Receives earnings" if they report one or more household members receiving income of $800 or more (qn18c). A respondent's household is designated "Receives public assistance" if they report one or more household members receiving food stamps (qn30a), TANF (qn31a), Refugee Cash Assistance (RCA) (qn32a), Supplemental Security Income (SSI) (q33a), or General Assistance (GA) (qn34a) or residing in public housing (qn38c) in the twelve months before survey administration. If both are true, households are designated "Receives both;" if neither is true, households are designated "Does not receive earnings or public assistance." If a respondent reports their household receiving either public assistance or earnings, and doesn't know or refuses to answer regarding the other, their household is designated either "Receives public assistance, but earnings missing" or "Receives earnings, but public assistance missing." If a respondent reports their household not receiving public assistance and doesn't know or refuses to answer regarding earnings, their household is designated "Doesn't receive public assistance, but earnings missing." Finally, if a respondent either doesn't know, refuses to answer, or has a missing value for both the earnings and the public assistance questions, their household is designated "Don't know and/or refused" for *ui_soi*.

ui_qn8a_annual: This variable reports estimated annual earnings from the individual's primary job. It was calculated by converting responses to qn8a (pre-tax earnings from primary job) to annual levels using responses to qn8b (basis on which *qn8a* was computed: weekly, bi-monthly, monthly, or annually). Individuals who answered "weekly" to question 8b were assumed to work fifty weeks in a year. Individuals who answered "bi-monthly" and "monthly" to question 8b were assumed to work twelve months in a year.

ui_qn10a_annual: This variable reports estimated annual earnings from the individual's secondary job. It was calculated by converting responses to qn10a

(pre-tax earnings from primary job) to annual levels using responses to *qn10b* (basis on which qn10a was computed: weekly, bi-monthly, monthly, or annually). Individuals who answered "weekly" to question 10b were assumed to work fifty weeks in a year. Individuals who answered "bi-monthly" and "monthly" to question 10b were assumed to work twelve months in a year.

ui_work: This variable reports individuals' present and past work status in the U.S.: working now, not working now but worked in past, not working now and never worked in past, not working now but unsure about working in past, not working now and refused regarding past, don't know for both, and refused for both. It was created using responses to qn5a and qn11a. Individuals are designated "Working now" if they report working at a job anytime the week before survey administration (qn5a). Individuals are designated "Not working now but worked in past" if they report not working at a job anytime the week before survey administration (qn5a) but working at some point since coming to the U.S. (qn11a). They are designated "Not working now and never worked in past" if they responded accordingly to qn5a and qn11a. They are designated "Not working now and unsure about working in past" or "Not working now and refused about past" if they reported not working anytime the week before survey administration (qn5a) and answered qn11a "don't know" or "refused," respectively. Respondents who either don't know or refuse to respond to both qn5a and qn11a are marked "Don't know and/or refused" for *ui_work*.

Administrative Variables

hhid: This number is used to identify which household each person lives in.

numppl: The number of people residing in each household. Respondents could list up to five household members including themselves.

language: The language of survey administration. For sampling and survey administration purposes, language data was borrowed from the RADS. Respondents could request that the survey be administered to them in a different language. The *language* variable reflects these changes.

cohort: For sampling and analysis purposes, respondents were divided into cohorts by year of arrival. The three cohort groups are refugees who arrived in fiscal years 2012 and 2013, fiscal years 2014 and 2015, and fiscal year 2016.

personid: This number uniquely identifies individuals in the dataset. The variable was created by combining *hhid* and *qn1a*.

respondent: This binary variable flags survey respondents as "1" and the household members they listed and reported on as "0." It is important to note that the survey respondents were the Principal Applicant (PA) whose refugee case is the basis for admission. This person is often the head of the household. The PAs provided self-reported data about themselves and proxy reported data for other household residents. ***To conduct household-level analysis it is recommended that you select only cases where this flag equals "1."***

Section 3: Dealing with Missing Respondent Data

Often when working with large federal data files such as the American Community Survey (ACS) or the Current Population Survey (CPS) it may seem as though most of the variables have no missing data. As with all surveys, however, some respondents either cannot answer or choose not to answer certain questions. The reason for most of the CPS variables having no missing data is this: when they are not obtained from the respondent, the answers are imputed by the Census Bureau through an elaborate imputation procedure.

Imputation did not occur for ASR questions. Therefore, most ASR questions have some missing data. For instance, if you add together the number of people who said yes or no for any yes/no question, that sum is almost always less than the total number of ASR respondents who were asked the question because of missing data. It is sometimes important to differentiate between two types of missing data: 1) data missing because the respondent does not provide a useable answer and 2) data missing by design because the respondent was purposely not asked the question (i.e., inapplicable).

The data file uses two codes to indicate when a respondent does not provide a useable response. When the answer to the ASR question is a "9" or repeated “9”s, this indicates that the respondent refused to answer the item. This could simply be a “9” or it could be “99”, “999”, “9999”, “99999” depending on the range of response options. For instance, the question that asks how well each person speaks English uses a “9” to indicate a refusal response while the question that asks about number of years of schooling uses a “99” to indicate a refusal. Similarly, when the answer to an ASR question is an “8” or “9”s ending in an “8”, this indicates that the respondent said that they don’t know the answer. Again, this could simply be an “8” or it could be “98”, “998”, “9998”, “99998” depending on the range of response options.

The data field is left empty or blank for variables that have missing data because the respondent did not get asked the question due to survey skip logic. For example, the 1,487 respondents who answered “No” to question 5a (“Did this person work at a job anytime last week?”) have an empty or blank response for questions 5b and 5c, since these are follow-up questions asked only of people who answered “Yes” to question 5a (i.e., reported working at a job in the previous week).

Usually researchers will exclude respondent missing data when calculating percentage estimates. This practice can be thought of as a form of pseudo-imputation—with the assumption that data missing from respondents would likely show a similar response pattern as the non-missing data.

 *To match the percentage estimates that are published in the Office of Refugee Resettlement Annual Report to Congress, you should exclude respondents' missing data.*

Information that is missing by design is typically excluded by researchers when producing percentage estimates. For some analyses, however, sometimes the items that are purposely not asked can be interpreted as having a value. For instance, question qn11a asks whether a person has ever worked since coming to the U.S. to stay. If you were to simply exclude all the missing data, you would conclude that the answer would be that 71.4% never worked since coming to the U.S. However, the answer you probably want to report is that only 31.5% of refugees never worked since coming to the U.S. This is because on an earlier question (qn5a) we learned that 1,607 people had a job last week and thus did not get asked about whether they have ever worked in the U.S. In this and many similar situations, it's important to carefully handle missing data.

 *Hence, when you have a question that has lots of missing data, consider checking the survey instrument (Appendix A) to see if respondents were not asked the question because of their response to a previous question.*

 *Whenever you are doing statistical analysis with variables that have missing data, make sure you either understand how the missing data is being treated or include in your program explicit instructions about how missing data will be handled. Given the large sample size for many of the ASR variables, you may not notice the unintentional impact that values of 8, 9, 98, or 99 may have on estimates such as means, medians, and regression coefficients.*

The next section of this report will show that when you exclude missing data, your weighted population estimates will no longer sum to the total refugee population and, therefore, will not produce accurate population estimates. The next section also explains, however, that there are acceptable procedures for producing reliable population estimates.

Section 4: Using the Survey Weights

Responses to ASR questions should be weighted to provide approximately unbiased aggregate estimates. The weights should be applied to all survey items in order to:

- Compensate for differential probabilities of selection for households and persons and
- Reduce biases occurring where non-respondents have different characteristics than respondents.

Household- and person-level analytic weights were developed for the 2017 ASR to allow for valid statistical estimates of the target refugee population. Both sets of weights are comprised of two components – a base weight reflecting the selection probability and an adjustment that corrects for differential nonresponse and aligns the population to known totals from the sampling frame (RADS universe file).

 Specifically, analytic weights incorporate:

- A ***base (sampling) weight*** which reflects the refugee household selection probability. The weight itself is simply the reciprocal of the probability of selection; because the sample allocations of each cohort were managed separately, the selection probabilities varied by the size of the population and the amount of sample released into the field;
- A ***post-stratification adjustment*** which corrects the sample for differential nonresponse[4] across cohort and demographic subgroups as well as aligning the sample to known population distributions taken from the RADS.

An important technical point is that the selection probability of persons in the ASR is the same as the household selection probability because the household-level respondent (i.e., the PA) served as a proxy for providing data on all eligible household residents.

 ***Household Analytic Weights.***

The household analytic weight was formed by taking the product of the base weight and post-stratification adjustment. The post-stratification adjustment was

[4] A separate nonresponse adjustment had been planned, but was not feasible. The reasons were the overwhelming nature of nonresponse stemming from being unable to locate subjects combined with a consistently high level of survey participation when subjects were contacted. There was virtually no useful information for nonresponse adjustment in the sample management system other than that which came from the RADS. Hence the post-stratification adjustment served both purposes of adjusting for nonresponse and aligning to population distributions.

created by applying an iterative proportional fitting algorithm called "raking" to the ASR household-level respondent data. This created weight adjustments that simultaneously align the sample to the refugee population distributions for several factors from the RADS data. Specific household-level post-stratification raking variables included:

- Year of entry
- Origin country collapsed to top 5 and all other
- Family size with 5+ collapsed into one category
- Voluntary agency collapsed to top 5 and all other
- U.S. state originally settled in collapsed to Census regions

Extreme weight adjustments, at both the top and bottom of the distribution of adjusted values, were trimmed to reduce the statistical variance associated with extreme weight values. Diagnostics comparing ASR weighted distributions to their corresponding RADS distributions were produced to verify that the final analytic household weight performed satisfactorily. Table 9 presents the results of that comparison.

 ***Person Analytic Weights.***
Similarly, the person-level analytic weight was formed by conducting the post-stratification raking procedure on a set of person-level RADS factors.[5] These included:

- Year of entry
- Origin country collapsed to top 5 and all other
- Family size at arrival with 5+ collapsed into one category
- Voluntary agency collapsed to top 5 and all other
- U.S. state originally settled in collapsed to Census regions
- Age at arrival collapsed into 5 categories [0-15, 16-24, 25-39, 40-54, and 55+]
- Gender
- Ethnicity collapsed to top 7 and all other
- Educational attainment collapsed into 6 categories (none/kindergarten, primary, intermediate, secondary, postsecondary, and unknown/missing)
- Language collapsed into top 5, unknown/missing, and all other.

[5] Since the focus of the survey is on persons 16 years of age or older, the person-level post-stratification weighting was done separately for persons 15 years of age or younger versus persons 16 years of age or older.

Table 9: 2017 ASR Household-level (PA) Distributions Using Analytic Weight Compared to RADS Household (PA) Distributions

Variable	Category	2017 ASR weighted	RADS principal applicant universe
Fiscal year of arrival	2012	18%	18%
	2013	21%	21%
	2014	21%	20%
	2015	19%	19%
	2016	21%	21%
Origin country	BURMA	22%	22%
	IRAQ	21%	21%
	BHUTAN	13%	13%
	SOMALIA	11%	11%
	DEM. REP. CONGO	7%	7%
	OTHER	26%	26%
Family size at arrival	1	51%	51%
	2	12%	12%
	3	12%	12%
	4	11%	11%
	5+	14%	14%
Region of placement	Northeast	15%	15%
	Midwest	27%	27%
	South	32%	32%
	West	25%	25%
Voluntary agency	US CONFERENCE OF CATHOLIC BISHOPS	25%	25%
	LUTHERAN IMMIGRATION AND REFUGEE SERVICE	14%	14%
	INTERNATIONAL RESCUE COMMITTEE	13%	13%
	US COMMITTEE FOR REFUGEES AND IMMIGRANTS	12%	12%
	CHURCH WORLD SERVICES	10%	10%
	OTHER	27%	27%

Table 10: 2017 ASR Person-Level Distributions of Person Aged 16+ Using Analytic Weight Compared to RADS Population Distributions

Variable	Category	2017 ASR weighted	RADS Person-level Universe
Fiscal year of arrival	2012	18%	18%
	2013	21%	21%
	2014	21%	20%
	2015	20%	19%
	2016	22%	21%
Origin country	IRAQ	22%	22%
	BURMA	20%	21%
	BHUTAN	14%	14%
	SOMALIA	10%	10%
	DEM. REP. CONGO	8%	8%
	OTHER	26%	26%
Family size at arrival	1	30%	30%
	2	13%	12%
	3	16%	15%
	4	15%	16%
	5+	27%	27%
Region of placement	Northeast	15%	15%
	Midwest	27%	27%
	South	32%	32%
	West	25%	25%
Voluntary agency	UNITED STATES CONFERENCE OF CATHOLIC BISHOPS	24%	25%
	LUTHERAN IMMIGRATION AND REFUGEE SERVICE	14%	14%
	INTERNATIONAL RESCUE COMMITTEE	12%	12%
	UNITED STATES COMMITTEE FOR REFUGEES AND IMMIGRANTS	13%	12%
	CHURCH WORLD SERVICES	10%	10%
	OTHER	27%	27%
Age at arrival	0-15	8%	9%
	16-24	25%	26%
	25-39	39%	39%
	40-54	18%	18%
	55+	10%	10%
Gender	Male	53%	53%
	Female	47%	47%

Ethnicity	ARAB	18%	17%
	LHOTSAMPA	14%	14%
	CHIN	8%	8%
	KAREN	6%	7%
	CUBAN	4%	4%
	CHALDEAN	4%	4%
	DAROD	3%	3%
	ALL OTHER	41%	43%
Language	Arabic	23%	22%
	Nepali	15%	14%
	Somali	10%	10%
	Sgaw Karen	6%	5%
	Spanish	5%	5%
	Other	42%	44%
Education	None/Kindergarten	1%	2%
	Primary	25%	26%
	Intermediate	15%	15%
	Secondary	29%	28%
	Postsecondary	15%	15%
	Unknown/Missing	15%	15%

As with the household analytic weight, extreme adjustments at both the top and bottom of the distribution of adjusted values were trimmed to reduce the statistical variance associated with extreme weight values. Diagnostics comparing person-level ASR weighted distributions to their corresponding RADS distributions were produced to verify that the final analytic household weight performed satisfactorily. Table 10 presents the results of that comparison.

Perhaps the most important task and one of the first tasks facing the data user will be determining whether you want to do person-level or household-level analysis.

For person-level analysis you would use the weight variable" Weight_person" or the weight variable "Weight_person_pop". These two person-level weight variables will produce the same estimates. However, when using the "Weight_person" variable the frequency counts will sum to ASR sample size of 4,111 and when using the "Weight_person_pop" variable the frequency counts will sum to the population of 353,078.

 *For household-level analysis, you need to filter the data file so that you have one observation per household. The easiest way to do this is to select only observations where the value of the "respondent" variable is equal to 1.*

After selecting 1,515 observations where the respondent variable equals 1, you would use the weight variable "Weight_household" or the weight variable "Weight_household_pop" to get household-level estimates. These two household-level weight variables will produce the same estimates. However, when using the "Weight_household" variable the frequency counts will sum to the ASR sample size of 1,515 and when using the "Weight_household_pop" variable the frequency counts will sum to the population of 146,599.

The data file also includes 20 replicate weights for each of the four survey weights on the data file (Weight_person, Weight_person_pop, Weight_household, Weight_household_pop). Replicate weights were created for each replicate sample to make it easier to estimate standard errors and confidence intervals which is covered in the section 5 of this user's guide.

The following table (Table 11) presents an example of basic descriptive analysis using the person-level sample and population main weights. It shows the unweighted and the person-level weighted estimates for the number of refugees (16 years old or older at time of survey administration and entered the U.S. as refugees between FY 2012 and FY 2016) that had English language instruction before coming to the United States. The shaded portion of the table shows unweighted estimates, while the non-shaded region of the table shows the person-level weighted estimates.

TABLE 11 English language instruction before coming to the United States (qn4c)			
Unweighted Frequency[6]			
Response Option	Frequency	Percent	Percent excluding missing data
1= NO	2,244	72.2	72.7
2= YES	843	27.1	27.3
8= DON'T KNOW	15	05	-
9=REFUSAL	7	0.2	-
Total	3,109	100.0	100.0
Weighted Frequency (using the person-level sample weight, Weight_person)			
Response Option	Frequency	Percent	Percent excluding missing data
1= NO	2,038	70.4	71.1
2= YES	831	28.7	28.9
8= DON'T KNOW	17	0.6	-
9=REFUSAL	8	0.3	-
Total	2,893	100.0	100.0
Weighted Frequency (using the person-level population weight, Weight_person_pop)			
Response Option	Frequency	Percent	Percent excluding missing data
1= NO	175,053	70.4	71.1
2= YES	71,333	28.7	28.9
8= DON'T KNOW	1,468	0.6	-
9=REFUSAL	654	0.3	-
Total	248,508	100.0	100.0

The weighted frequency using the sample person-level weight sums to 2,893 rather than the unweighted sample size of 3,109. This means that children under 16 years of age are overrepresented in the unweighted sample relative to refugees 16 years old or older. The person-level weight adjusts so that all refugees 16 years old or older at time of survey administration that entered the country between FY 2012 and FY 2016 have the same probability of being in the sample. Note that the person-level sample and population weights will generate the same percent estimates, but the frequency counts for the person-level population weight sums to the overall estimate population of refugees 16 years old or older at time of survey administration who entered the U.S. as refugees between FY 2012 and FY 2016.

The following table (Table 12) shows the unweighted and the household-level weighted estimates for the number of refugee households in which one or more persons received food stamps in the past 12 months. The shaded portion of the

[6] These frequencies were tabulated excluding individuals who were not found to be eligible refugees and therefore do not have a person weight.

table shows unweighted estimates, while the non-shaded region of the table shows the household-level weighted estimates.

TABLE 12
The number of refugee households in which
one or more persons received food stamps in the past 12 months (qn30a)

Unweighted Frequency (filtering by respondent=1).

Response Option	Frequency	Percent	Percent excluding missing data
1= YES	563	37.2	37.4
2= NO	944	62.3	62.6
8= DON'T KNOW	7	0.5	-
9=REFUSAL	1	0.1	-
Total	1,515	100.0	100.0

Weighted Frequency (using the household-level sample weight, Weight_household, and filtering the data by respondent=1).

Response Option	Frequency	Percent	Percent excluding missing data
1= YES	660	43.6	43.9
2= NO	845	55.8	56.1
8= DON'T KNOW	9	0.6	-
9=REFUSAL	1	0.1	-
Total	1,515	100.0	100.0

Weighted Frequency (using the household-level population weight, Weight_household_pop, and filtering the data by respondent=1).

Response Option	Frequency	Percent	Percent excluding missing data
1= YES	63,855	43.6	43.9
2= NO	81,756	55.8	56.1
8= DON'T KNOW	911	0.6	-
9=REFUSAL	77	0.1	-
Total	146,599	100.0	100.0

By checking the frequency count, you usually can tell that you are looking at a population weighted estimate. A population weighted frequency count will have much larger numbers compared with the sample size.

Although a large frequency count tends to indicate a weighted population estimate, the statistical output is usually not helpful in determining whether the correct survey weight was applied. Table 13 demonstrates how similar the two sets of estimates are when the ***wrong*** weight is applied. The shaded portion of the table shows the estimate that would have resulted by incorrectly using the household-level population weight (Weight_household_pop) instead of the person-level population weight (Weight_person_pop). The percentage estimates are so similar that even an experienced researcher may be unable to tell just from the statistical output whether the appropriate weight was used.

TABLE 13 "Within the past 12 months, has this person attended any job training program?" (qn24a)			
Incorrectly Weighted Frequency Using Household-level population weight			
Response Option	Frequency	Percent	Percent excluding missing data
1= YES	123,470	84.2	84.7
2= NO	22,292	15.2	15.3
8= DON'T KNOW	795	0.5	-
9=REFUSAL	43	0.0	-
Total	146,599	100.0	100.0
Correctly Weighted Frequency Using Person-level population weight			
Response Option	Frequency	Percent	Percent excluding missing data
1= YES	214,499	86.3	88.5
2= NO	27,878	11.2	11.5
8= DON'T KNOW	5,235	2.1	-
9=REFUSAL	896	0.4	-
Total	248,508	100.0	100.0

Based on the State Department admissions report, there were 353,078 refugees (of all ages) that entered the U.S. in FY 2012-2016, and 248,508 of these would be 16 or older at the time of the survey. The person-level population weights can be used to estimate answers to survey questions like how many of these approximately 248,500 refugees 16 or older received a benefit or were working last week. When doing population estimates, however, you must be careful of how you handle missing data. Missing data typically occur when a person refuses to answer a question or does not know the answer to the question being asked. These situations are usually lumped together and classified as "missing data." Refer to the previous section of the guide for more information on how missing data for the ASR variables have been coded.

If you don't exclude missing data from frequency estimates, your population counts will total to less than the overall population of 248,508 refugees 16 or older. For instance, in our table 11 example, the estimated total number of refugees who did not have English language instruction before coming to the United States equaled 175,053, or 70.4% of the refugee population. However, these estimates do not treat as missing those respondents who answered "don't know" or refused to answer, so they don't sum to 100% of the refugee population when combined with refugees who **did** have English instruction before coming to the United States. To avoid this issue, most researchers exclude missing data when reporting estimates. In this case, that means reporting that 71.1% of the refugees 16 or older did not have English language instruction before coming to the United States (the "percent excluding missing data" column in table 11).

If you decide to exclude the missing data, then a more accurate population estimate will be obtained by multiplying the percentage that excludes missing data by the total population. For instance, when we exclude the missing data from the previous example, we see that 71.1% of the refugees 16 or older did not have English language instruction before coming to the United States. This proportion translates to approximately 176,689 (.711 x 248,508 refugees) refugees 16 or older that did not have English language instruction before coming to the United States as opposed to the estimate of 175,053 shown in table 11. Again, this difference occurs because the population estimates in table 10 do not adjust for the missing data.

Section 5: Procedures for Estimating Standard Errors

The sample of households and persons surveyed for the 2017 Annual Survey of Refuges (ASR) is just one of many possible samples that could have been drawn. Sampling error refers to error in survey estimates that arise from the fact that estimates are based on a sample of observations rather than the whole population. This form of error is usually expressed in terms of the sampling variance or standard error of an estimate, which is simply the square root of the sampling variance. Standard errors are required to calculate margins of error (i.e., the half width of a confidence interval) or to conduct hypothesis tests or tests of statistical significance. A clear presentation of estimates from a survey or hypothesis test should include measures of uncertainty associated with using a sample for inference, as opposed to using the entire population.

This section explains the process of obtaining standard errors for the 2017 ASR estimates. The 2017 ASR sample and respondents are subsets of all refugees who entered the country between fiscal years 2012 and 2016. Although survey estimates obtained from the default options in most statistical packages will be correct, the associated standard error estimates will often understate the true standard errors because they do not account for the weighting, clustering of persons within households and survey design (e.g., oversampling and stratification).

Stratification generally leads to a gain in efficiency over simple random sampling. On the other hand, clustering usually leads to deterioration in efficiency. This latter effect arises because of the positive intra-cluster correlation (i.e., similarity) among the subjects within the sampling clusters. For example, respondents from the same household are expected to have a higher likelihood of having the same ethnicity, religion, and country of origin than respondents selected at random from the list of all refugees that arrived during the target period. The cluster effect is larger for larger households because the survey sampled every eligible refugee from the same household, and this clustering effect increases the variance over what would pertain in a simple random sampling of refugees.

To determine the total effect of any complex survey design on the sampling variance, users must first calculate the variance associated with an estimate assuming a complex sample design. Then users calculate the variance expected from a simple random sample design. The ratio of the complex variance estimate over the variance associated with a simple design is what is called the design effect, often referred to as the DEFF, and it measures the overall efficiency of the survey weights and sample design.

In a wide range of situations, the adjusted standard error of a statistic should be calculated by multiplying the usual formula by the square root of the DEFF. Thus, the formula for computing the 95% confidence interval around a percentage is:

$$\hat{p} \pm \left(deft \times 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

where *p* is the sample estimate, *n* is the unweighted number of sample cases in the group being considered, and *deft* is the square root of DEFF.

The remainder of this section discusses how to use the replicate weights that are included on the data file to estimate the overall average design effect and to estimate design effect separately for each estimate. Both household-level and person-level replicate weights are included on the data file and can be used to obtain standard errors reflecting the complexity of the ASR sample design. However, for researchers who may not have access to the necessary computer hardware and software or technical ability to use these replicate weights to calculate standard errors appropriately, you should at least use the overall estimated average design effect to obtain approximate standard errors for survey estimates.

The overall square root of the average design effect for household-level analysis is 1.13. For person-level analysis that includes persons of all ages, the overall square root of the average design effect it is 1.29. For persons 16 or older, the square root of the average design effect is 1.16.

 *Multiplying your standard error estimates by the square root of the overall design effect will provide much more appropriate standard error estimates associated with your ASR estimates than incorrectly using the simple random sample estimates of variance, e.g., using [p x (1-p)] /n as the variance of a proportion p.*

 *Still, it is important to keep in mind that each survey estimate has its own design effect. Therefore, the design effect for receiving food stamps may be higher or lower for, say, families with children compared to families without children or for any other subgroup of the population. If getting more precise standard estimates is a concern, then follow the instructions in the remainder of this section on how to use the replicate weights to estimate standard errors.*

We now discuss how to calculate standard errors for the ASR estimates using the 80 replicate weights that are included on the 2017 ASR data files. Table 14 shows the names of the 20 replicate weights for each of the four main survey weights on the data file.

Table 14

	Person-level sample weights	Person-level population weights	Household-level sample weights	Household-level populations weights
Main Weight Variable	Weight_person	Weight_person_pop	Weight_household	Weight_household_pop
Replicate Weight Variables	Weight_person_R1 through Weight_person_R20	Weight_person_pop_R1 through Weight_person_pop_R20	Weight_household_R1 through Weight_household_R20	Weight_household_pop_R1 through Weight_household_pop_R20

The basic idea behind replication is to draw subsamples from the sample, compute the estimate from each of the subsamples, and estimate the variance from the variability of the subsample estimates. Specifically, subsamples of the original full sample are selected to calculate subsample estimates of a parameter for which a full-sample estimate of interest has been generated. The variability of these subsample estimates around the estimate for the full sample provides an estimate of the standard error of the estimate. The subsamples are called replicates and the estimates from the subsamples are called replicate estimates.

Although the logic behind using replicate weights is not unduly complicated, it can be compute-intensive to produce standard errors using the replicate weights. To use the replicate weights, users can either use specialized software designed to make use of replicate weights when generating standard errors— examples include SUDAAN and WesVar— or use specialized advanced sampling modules in software such as Stata, SAS, or SPPS. Below is an example of using Stata survey commands to estimate means:

Survey set:
svyset _n [iweight=Weight_person] , jkrweight(Weight_person_R1 Weight_person_R2 Weight_person_R3 Weight_person_R4 Weight_person_R5 Weight_person_R6 Weight_person_R7 Weight_person_R8 Weight_person_R9 Weight_person_R10 Weight_person_R11 Weight_person_R12 Weight_person_R13 Weight_person_R14 Weight_person_R15 Weight_person_R16 Weight_person_R17 Weight_person_R18 Weight_person_R19 Weight_person_R20) vce(linearized)

Estimate mean w/ SE:
svy, vce(jackknife): mean varname

Example:
Below, we calculate the standard error for the mean of *numppl* (number of people in the household) in Stata.

Survey set:
svyset _n [iweight=Weight_person] , jkrweight(Weight_person_R1 Weight_person_R2 Weight_person_R3 Weight_person_R4 Weight_person_R5 Weight_person_R6 Weight_person_R7 Weight_person_R8 Weight_person_R9 Weight_person_R10

Weight_person_R11 Weight_person_R12 Weight_person_R13 Weight_person_R14 Weight_person_R15 Weight_person_R16 Weight_person_R17 Weight_person_R18 Weight_person_R19 Weight_person_R20) vce(linearized)

Estimate mean w/ SE:
svy, vce(jackknife): mean numppl

Output:

```
Survey: Mean estimation

Number of strata =       1        Number of obs   =      4,111
                                  Population size = 4,111.0037
                                  Replications    =         20
                                  Design df       =         19
```

	Mean	Jackknife Std. Err.	[95% Conf.	Interval]
numppl	3.971903	.0345228	3.899646	4.04416

Section 6: Comparing 2017 ASR to Earlier ASR Estimates

The comparison of the ASR estimates over time is something that many researchers often do. Although the 2017 ASR is not a longitudinal study, the estimates can be compared with estimates from earlier ASR studies. As with any survey, there are limits to how much a change can be considered "real" and not reflect larger differences in the sampled population or in the methodologies used for each round of data collection. This section of the user's guide will describe key factors that should be considered when comparing the 2017 ASR with earlier ASR studies.

 *Due to the considerable differences in survey methodologies, researchers should be cautious when comparing 2017 ASR estimates with ASR estimates prior to 2016. Hence, the discussion and considerations covered in this section are applicable only to comparisons done between the 2017 ASR and the 2016 ASR.*

To compare 2017 estimates with prior estimates, you need to obtain a copy of earlier questionnaires and compare the wording of those questions you plan to analyze with the 2017 wording found in Appendix A of this user's guide. Differences in question wording do not necessarily mean that you cannot compare changes in estimates over time. Such changes may have been necessary to improve the questions. Admittedly, the impact of wording changes is a matter for subjective judgment. Even if the perceived impact is minor, it is generally good practice to acknowledge in an endnote or footnote when there are wording differences.

To understand the simplicity of estimating the significance of changes over time, consider estimating a proportion or count at time t — say, t θ_t. Let v(θ_t) be its estimated variance (the square of the standard error). The estimated change between times t1 and t2 for this proportion or count is $\Delta = \theta_{t1} - \theta_{t2.}$ The variance of the difference is the sum of the variances for the two-time periods, which is $v\Delta = (v\theta_{t1\_+} v\theta_{t2})$. The two variances on the right side of the equation should be computed separately. To get the standard error of the differences between the two estimates you would then take the square root of $v\Delta$. If the difference between the two estimates is greater than 1.96 times the standard error of the differences, then you can say with 95% certainty that the differences between the different ASR estimates are significant.

Table 15 provides a real example of how you would go about determining whether a change in an estimate is significant. In this example, we see that the percentage of households with someone receiving cash assistance increases by 2.7 percentage points from the 2016 ASR to the 2017 ASR. Is that a statistically significant change?

The first step we need to take to answer this question is to sum the adjusted variances of the two estimates. The sum would be equal to 0.44 (fourth column) The next step is to estimate the standard error of the difference, which is the square root of the adjusted variance, or 0.66. Finally, we can build a confidence interval by multiplying the adjusted standard error by 1.96 (this step gives you a 95% confidence interval), and then adding and subtracting that number to and from the 2.8 percentage point change that occurred between 2006 and 2007.

The answer to our hypothetical question is yes. We are at least 95% confident that there was an increase in cash assistance from 2006 to 2007, as zero (no change) is not within the confidence interval (1.4 to 4.0).

TABLE 15

Testing to see if the change in households with someone receiving cash assistance between the 2016 and 2017 ASRs was statistically significant at the 95% confidence interval

	Percent	Variance	Design Effect	Adjusted Variance	Adjusted Standard Error	95% Confidence Interval
2016	25.2	.19	1.13	.21	.46	24.3 to 26.1
2017	27.9	.20	1.14	.23	.53	26.9 to 28.9
(2017-2016)	2.7		NA	.44	.66	1.4 to 4.0

When comparing responses over time, there are a few other things worth thinking about besides question wording and sampling. For instance, even though the survey weights include a nonresponse adjustment, differences in ASR response rates may explain small changes in the estimates.

Sample sizes for the ASR studies are not always sufficient for producing some estimates. For example, it becomes even more difficult to look at differences by region or country of origin when participation rates are quite small, as in the case of Refugee Cash Assistance (RCA) or Temporary Assistance for Needy Families (TANF). If we were to pool data across ASR years, then we would be able to increase the number of observations. However, to analyze combined responses to the same question across multiple years (pooled data) you need to normalize the population weight variable from each ASR year being pooled. This is effectively the same as averaging the population weight variable across years.