UNITED STATES DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. Census Bureau
Washington, DC 20233-0001

July 24, 2013

2013 AMERICAN COMMUNITY SURVEY RESEARCH AND EVALUATION REPORT MEMORANDUM SERIES #ACS13-RER-09

MEMORANDUM FOR	ACS Research and Evaluation Advisory Group (email list "ACSO RE Advisory Group List")
From:	James B Treat **(Signed 7/26/2013)** Chief, American Community Survey Office
Prepared by:	Samantha Fish Special Studies Staff American Community Survey Office
Subject:	Evaluation of the Length of Write-In Entries and the Potential Effect of Truncation of Long Responses

Attached is the final American Community Survey Research and Evaluation report for the Evaluation of the Length of Write-In Entries and the Potential Effect of Truncation of Long Responses.

This report documents the maximum field lengths specific to each open response question in the 2009 ACS and measures rates of truncated responses in the mail mode. We assess the 2009 ACS data because we originally conducted this analysis in 2010 but did not officially report the results. The Internet data collection mode did not exist at that time, so it is not discussed in the report. These results have helped/ will help managers evaluate what, if any, changes may be needed in the data capture and processing methods to improve data quality.

If you have any questions about this report, please contact David Raglin (301-763-4226) or Samantha Fish (301-763-7542).

Attachment
cc:
ACS Research and Evaluation Workgroup (email list "ACSO RE Workgroup List")
Kenneth Dawson, ACSO
Dameka Reese, ACSO
Ellen Wilson, SEHSD

JULY 24, 2013

Evaluation of the Length of Write-In Entries and the Potential Effect of Truncation of Long Responses

FINAL REPORT

Samantha Fish
American Community Survey Office

United States Census Bureau

Table of Contents

Introduction 1

Research Questions 1

Background 2

Methodology 2

Limitations 3

Results 3

1. What are the maximum response lengths for open-ended survey items in the 2009 ACS? 3

2. How often do open-ended responses in the mail mode exceed their maximum character lengths for data processing? 5

3. How long are the mail open-ended responses that exceed their maximum character lengths for data processing? 7

4. How often do open-ended responses in the automated modes equal their maximum character length in data collection? 10

5. Would expanding the maximum character lengths in data collection or subsequent data processing for certain survey items capture more meaningful data? 12

Conclusions 12

Appendix A 14

Introduction

Starting in mid-2007, the American Community Survey (ACS) began using Optical Mark Recognition (OMR) and key-from-image (KFI) software to capture data from returned paper questionnaires. Keyers recorded respondent data directly from scanned computer images of the returned forms instead of finding and flipping through them physically. The adoption of this new software provided ACS the opportunity to capture more characters for write-in entries on the mail forms. For the first time, keyers could enter respondent write-ins up to a very large maximum length.

However, many subsequent coding operations truncate these data before coders view them because the coding software requires certain input lengths, which we cannot easily change.[1] To assess whether we should spend time and resources to make changes, the ACS archived the "before and after" truncation values for all write-in entries. The ACS program planned to use these files to measure how often truncation occurs. This would help management decide if expanding field lengths for write-in entries in the coding software is worth the cost to make the changes.

Additionally, the ACS collects data in the Computer Assisted Telephone Interviewing (CATI) and Computer Assisted Personal Interviewing (CAPI) modes. CATI and CAPI interviewers use computer instruments to collect respondent data, which have various field sizes for open responses. Interviewers may abbreviate input into the instrument when the maximum field length is too small for a response. Although no truncation occurs in the subsequent coding of these responses, the ACS seeks to find out how often CATI and CAPI open responses meet maximum input lengths in the CATI and CAPI instruments.

This report documents the maximum field lengths specific to each open response question in the 2009 ACS, measures rates of truncation in the mail mode, and summarizes the distributions of response lengths in the mail and automated modes. Results of this research help evaluate what, if any, changes may be needed in the data capture and processing methods to improve data quality.

Research Questions

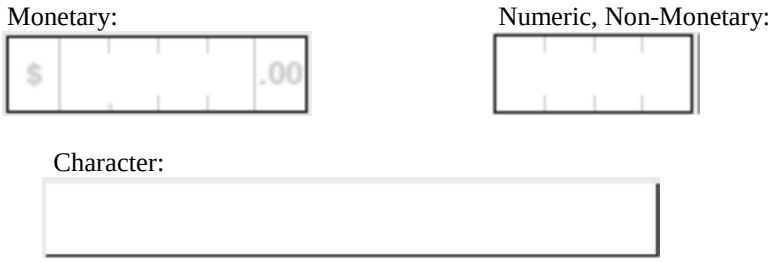
1. What are the maximum response lengths for open-ended survey items in the 2009 ACS?
2. How often do open-ended responses in the mail mode exceed their maximum character lengths for data processing?
3. How long are the mail open-ended responses that exceed their maximum character lengths for data processing?
4. How often do open-ended responses in the automated modes equal their maximum character length in data collection?
5. Would expanding the maximum character lengths in data collection or subsequent data processing for certain survey items capture more meaningful data?

[1] Many of these programs are written in old programming languages with the input lengths hard-coded in the code.

Background

ACS mail questionnaire responses are mostly checkbox responses, which the National Processing Center records using OMR technology. However, some questions require written responses. On the 2009 mail questionnaire, 26 items required a character response and 21 items required a monetary response. Figure 1 displays the write-in boxes for different types of questions.

Figure 1. Examples of Write-In Boxes on the 2009 ACS Questionnaire
Source: Form ACS-1(2009)KFI



Numeric answer boxes (monetary and non-monetary) segment for single digits while the character answer boxes are unsegmented. Specifically for monetary write-ins, the boxes come with a dollar sign in front and a ".00" at the end indicating that the response should be a whole dollar amount. All write-in response fields are positioned on a green background, which subtly deters respondents from writing outside of the answer box. To view the full 2009 ACS English questionnaire, please visit the Questionnaire Archive on the ACS Homepage at www.census.gov/acs/www/methodology/questionnaire_archive/.

We document research using the 2009 ACS data because we originally conducted this analysis in 2010 but we did not officially report on the results until now. We expect these results are similar to the recent rates of truncation in the mail and automated modes. Although the ACS introduced an Internet mode and a few new questions to the survey in 2013, only two new items require write-in responses (the Computer and Internet questions). The ACS plans to assess truncation rates for the Internet mode in a separate report.

Methodology

We assess character and monetary open-ended responses from the 2009 ACS English questionnaire using raw, response data from the 2009 panels (this differs slightly from the universe used to produce the 2009 estimates). The daily keying files were used for the analysis of returned mail forms, and the data capture file (DCF) was used for the analysis of responses from the automated modes. We chose to use the DCF instead of the actual CATI and CAPI output files because the raw output files required complex, custom programming to assemble them into a useable database.

We define "long responses" as strings of characters that exceed their maximum processing length. Long responses exist only in the mail mode and ACS processing truncates them prior to coding. The rate of truncation is the number of long responses for an open response item divided by the total number of responses for that item and multiplied by 100. We count all responses regardless of the respondent's eligibility for the question. We produce unweighted truncation rates to measure the occurrence of long responses among mail data processing. These rates represent only the incidence of truncation in mail data processing.

Limitations

This report summarizes raw ACS data, which may include multiple returns for a sample address. Multiple returns result when a sample address responds to both the first mailing and second mailings or we conduct CATI or CAPI and receive a late mail return. Generally, interviewers and field representatives try to obtain interviews even if the respondent says he/she returned the mail questionnaire because the interview may be a more complete return and sometimes the National Processing Center never receives the paper questionnaire. More than 98 percent of the 2009 responding addresses consisted of single returns, so this is not a major limitation.

This assessment covers all survey items with character and monetary open-ended responses except for those corresponding to the first and last names of household members in the basic and detailed sections. We do not assess these items individually due to dataset limitations; however, long response data regarding the respondent's first and last names are available and we expect that the lengths of responses to these fields should be similar to the lengths of responses in the first and last name fields in other sections of the survey. Thus, please use the respondent first and last name field results as a proxy for other first and last name field results.

Results

1. What are the maximum response lengths for open-ended survey items in the 2009 ACS?

The maximum character length for all character and monetary variables collected in the mail mode at the National Processing Center is currently 255. The ACS set this maximum length to be the maximum the system could accommodate so we could measure rates of truncation. For the automated modes, CATI and CAPI, the instruments have unique maximum lengths for each item. The specific lengths for these items are contained in Table 1.

After data capture, the ACS sends most open-ended items to automated and/or manual coding. Each coding program is item specific and responses sent to that operation must meet the maximum length requirements for the program. For items that do not require coding, the character strings are also truncated and numeric values are rounded to fit a specified length (for example, an entry of $36,000 for a 4 character field would become $9,999). The last column of Table 1 documents these length requirements, the Headquarters' expected lengths.

Table 1. Maximum Field Lengths for ACS Data Collection and Subsequent Processing
Source: 2009 Housing Unit Key From Image Matrix & the 2009 Automated Instrument Specifications

Character Variables		Maximum Length		
Description	**Variable Name**	**Mail Mode**	**Automated Modes**	**Headquarters' Expected Length**
Ancestry	ANCW	255	40	40
Field of degree write-in	FODW	255	75	75
Health insurance	HINSW	255	30	30
Hispanic origin write-in	HISW	255	30	30
First industry write-in	INW2	255	60	60
Second industry write-in	INW3	255	60	60
Language spoken at home	LANW	255	20	20
Migration foreign country	MGW1	255	30	30
Migration address	MGW2	255	50	53
Migration city	MGW3	255	20	20
Migration county	MGW4	255	20	20
Migration state	MGW5	255	2*	20
First occupation write-in	OCW1	255	60	60
Second occupation write-in	OCW2	255	60	60
Place of birth US	PBW2	255	2*	20
Place of birth outside US	PBW3	255	24	30
Place of work address	PWW1	255	53	53
Place of work city	PWW2	255	20	20
Place of work county	PWW4	255	20	20
Place of work state	PWW5	255	20	20
American Indian or Alaska Native Race	RCW1	255	30	30
Other Asian Race **or** Pacific Islander (MAIL); Other Race (AUTOMATED)	RCW2	255	30	30
Other Race	RCW3	255	30	30
Other Pacific Islander (AUTOMATED only)	RCW4	255	30	N/A**
Respondent first name (MAIL); Respondent whole name (AUTOMATED)	RFN	255	13	13
Respondent last name	RLN	255	20	20

Monetary Variables		Maximum Length		
Description	**Variable Name**	**Mail Mode**	**Automated Modes**	**Headquarters' Expected Length**
Monthly condominium fee	CON	255	4	4
Monthly electricity cost	ELE	255	4	4
Yearly other fuel cost	FUL	255	4	4
Monthly gas cost	GAS	255	4	4
Monthly insurance payment	INS	255	4	4
Interest Income	INT	255	6	6***
Yearly mobile home costs	MH	255	5	5
Monthly mortgage payment	MRG	255	5	5
Other income amount	OI	255	6	6
Publis assistance recipiency	PA	255	5	5
Retirement Income	RET	255	6	6
Monthly rent	RNT	255	5	5
Self-employment income	SEM	255	6	6***
Monthly other mortgage payments	SM	255	5	5
Social Security or Railroad Retirement income	SS	255	5	5
Supplemental security income	SSI	255	5	5
Yearly real estate tax	TAX	255	5	5
Total income	TI	255	7	7***
Property value	VAL	255	7	7
Wages/ salary income	WAG	255	6	6
Yearly water and sewer cost	WAT	255	4	4

Subsequent data processing limits both the mail mode and automated mode responses by the headquarters' expected field length. Since the mail field lengths are dramatically larger (by design), ACS truncates some mail responses in order to process them. The automated modes, however, have item field lengths very close to their respective Headquarters' expected field lengths, with a few exceptions: the Migration state, Place of birth US state, Migration address, Place of birth outside US, and Other Pacific Islander Race items.

First, the migration state (MGW5) and place of birth state (PBW2) variables are given a two-character, state code in automated instrument, but are translated into the full state name before being sent to coding; thus, this is not a problem. Next, the migration address (MGW2) and place of birth outside of the US (PBW3) items have instrument lengths that are 3 characters and 6 characters shorter than their respective maximum headquarters' expected lengths. Although this is inconsistent, it may not be a problem either because interviewers can abbreviate responses that are longer than the instrument field size.

The last variable on the list, the Other Pacific Islander Race (RCW4), does pose a problem in theory, but it has not been a problem in practice. ACS data processing concatenates the response to RCW4 to the end of the Other Race (RCW3) write-in. Because of the concatenation, the final concatenated string may be longer than the maximum length for coding. Thus, the ACS may truncate data collected in RCW4; however, this occurred only nine times from 2008 to 2010. See Appendix A for details. We suggest correcting this problem when the DCF processing is revised.

2. How often do open-ended responses in the mail mode exceed their maximum character lengths for data processing?

Table 2 shows the rates and frequencies of truncation by item in the mail mode for the 2009 panels. The frequencies state how often truncation occurred in the mail mode, while the rates show the proportion of mail responses that were truncated. We sort the variables in the table based on the descending truncation rate.

There are a few items with truncation rates above 1 percent—these are all survey questions requiring a character value response. The truncation rates for monetary variables were all less than 0.20 percent. The ACS truncates about 2.9 percent of American Indian or Alaska Native Race write-ins (RCW1), 1.1 percent of the Asian race or Pacific Islander race write-ins (RCW2), 1.1 percent of the Language Spoken at Home write-ins (LANW), and 1.1 percent of the Ancestry write-ins (ANCW). With the exception of ANCW, all items had fewer than 4,300 long responses in the 2009 ACS. The ANCW item had over 26,000 long responses. In general, these results suggest that there are low rates of truncation and a low volume of long responses in mail data collection.

Table 2. Mail Mode Rates of Truncation, Unweighted
Source: 2009 American Community Survey Sample, Mail Keyed Data

Character Variables

Description	Variable Name	Long Responses	Total Responses	Truncation Rate (%)
American Indian or Alaska Native race	RCW1	1,160	40,196	2.89
Other Asian race or Pacific Islander	RCW2	844	75,705	1.11
Language spoken at home	LANW	3,980	357,978	1.11
Ancestry	ANCW	26,407	2,453,431	1.08
Hispanic origin	HISW	377	79,548	0.47
Migration foreign country	MGW1	32	8,231	0.39
Respondent first name	RFN	3,750	1,266,437	0.30
Second occupation	OCW2	4,252	1,576,772	0.27
Other race	RCW3	70	26,680	0.26
Place of work city	PWW2	3,010	1,266,735	0.24
Place of work county	PWW4	1,370	1,209,284	0.11
Migration city	MGW3	396	360,593	0.11
Place of work address	PWW1	1,130	1,204,163	0.09
Place of birth outside US	PBW3	203	270,478	0.08
First industry	INW2	801	1,551,720	0.05
Second industry	INW3	721	1,615,667	0.04
Migration county	MGW4	104	235,596	0.04
First occupation	OCW1	598	1,643,155	0.04
Respondent last name	RLN	348	1,268,571	0.03
Place of birth US	PBW2	530	2,313,765	0.02
Migration address	MGW2	79	351,838	0.02
Place of work state	PWW5	60	1,257,834	<0.01
Migration state	MGW5	3	350,960	< 0.01
Field of degree	FODW	0	693,485	< 0.01
Health insurance	HINSW	0	153,116	< 0.01

Monetary Variables

Description	Variable Name	Long Responses	Total Responses	Truncation Rate (%)
Self-employment income	SEM	351	213,826	0.16
Monthly insurance payment	INS	1,264	775,991	0.16
Social Security or Railroad Retirement income	SS	529	502,126	0.11
Supplemental security income	SSI	102	108,673	0.09
Yealy real estate tax	TAX	743	856,110	0.09
Yearly other fuel cost	FUL	226	348,659	0.06
Wages/ salary income	WAG	823	1,412,591	0.06
Interest Income	INT	260	500,309	0.05
Monthly condomium fee	CON	42	84,848	0.05
Yearly mobile home costs	MH	38	83,407	0.05
Yearly water and sewer costs	WAT	375	854,697	0.04
Monthly other mortgage payments	SM	77	183,037	0.04
Property value	VAL	346	907,448	0.04
Public assistance recipiency	PA	13	68,401	0.02
Monthly mortgage payment	MRG	96	555,127	0.02
Retirement income	RET	45	316,994	0.01
Monthly electricity cost	ELE	137	1,178,023	0.01
Monthly gas cost	GAS	84	810,254	0.01
Other income amount	OI	17	168,943	0.01
Monthly rent	RNT	21	311,387	0.01
Total income	TI	116	1,850,264	0.01

3. How long are the mail open-ended responses that exceed their maximum character lengths for data processing?

Table 3 shows, for each open-ended survey question in the mail mode, the distribution of its long response lengths. Recall that we define long responses as the responses to open-ended survey questions that exceed the item's maximum allowable response length. The mean length and standard deviation of long response lengths are also given. Note that the 50th percentile length is the median length and the 100th percentile is the maximum length. We sort the table by rate of truncation and shade items with truncation rates above 1.0 percent.

Since the previous section identified four items in the mail mode that have truncation rates above 1.0 percent, we report on long response lengths for these items. These variables are the ancestry write-in (ANCW), the language spoken at home write-in (LANW), the American Indian or Alaska Native Race write-in (RCW1), and the Other Asian Race or Pacific Islander Race write-in (RCW2). Table 3 shows that the median long response length for LANW is 2 characters longer than the expected length, but for RCW1, RCW2, and ANCW the median long response length is about 6 characters longer than the limit. This suggests that mail processing may truncate the end of a word for LANW and, for RCW1, RCW2, and ANCW it may truncate a whole word or the end of a word.

Although mail processing truncates data, it is unknown if the truncated characters would make a difference in data quality if included. To judge this, the ACS would need to research how often response codes and/or the number of codes that coders assign would change if the coders instead coded from the full, pre-truncated responses. Research question 5 touches on this point.

Table 3. Mail Mode Distribution of Long Response Lengths, Unweighted

Source: 2009 American Community Survey Sample, Mail Keyed Data

Character Variables

Description	Variable Name	Headquarters' Expected Length	Number of Long Responses	Percentile Lengths of Mail Long Responses							Mean	Standard Deviation	Truncation Rate (%)
				25th	50th	75th	90th	95th	99th	100th			
American Indian or Alaska Native race	RCW1	30	1,160	33	36	41	46	50	62	89	37.7	7.0	2.89
Other Asian race or Pacific Islander	RCW2	30	844	33	36	41	49	57	68	110	38.5	9.4	1.11
Language spoken at home	LANW	20	3,980	22	22	26	31	34	47	70	24.9	5.3	1.11
Ancestry	ANCW	40	26,407	43	46	51	58	63	75	154	48.2	7.5	1.08
Hispanic origin	HISW	30	377	32	35	40	46	51	57	93	37.2	7.6	0.47
Migration foreign country	MGW1	30	32	33	39	51	54	55	55	55	41.0	8.5	0.39
Respondent first name	RFN	13	3,750	15	16	18	20	23	34	77	17.2	4.6	0.30
Second occupation	OCW2	60	4,252	64	69	80	98	111	151	255	75.8	19.5	0.27
Other race	RCW3	30	70	33	34	42	136	136	136	136	46.1	30.7	0.26
Place of work city	PWW2	20	3,010	22	24	29	37	44	61	84	27.1	8.4	0.24
Place of work county	PWW4	20	1,370	23	25	28	36	43	55	82	27.1	7.7	0.11
Migration city	MGW3	20	396	22	23	25	29	32	37	39	24.1	3.7	0.11
Place of work address	PWW1	53	1,130	56	60	67	78	85	111	134	63.5	11.1	0.09
Place of birth outside US	PBW3	30	203	32	33	37	45	56	72	72	36.0	7.9	0.08
First industry	INW2	60	801	63	68	75	86	93	116	140	71.4	11.7	0.05
Second industry	INW3	60	721	64	70	80	96	108	178	255	76.1	22.4	0.04
Migration county	MGW4	20	104	22	24	25	28	28	40	40	24.0	3.1	0.04
First occupation	OCW1	60	598	64	70	83	101	116	180	197	77.1	20.8	0.04
Respondent last name	RLN	20	348	22	24	29	34	39	48	68	26.3	6.7	0.03
Place of birth US	PBW2	20	530	22	24	28	33	39	62	81	26.4	7.4	0.02
Migration address	MGW2	53	79	55	58	63	71	79	83	83	60.7	7.1	0.02
Place of work state	PWW5	20	60	23	26	33	40	44	50	50	28.5	7.4	<0.01
Migration state	MGW5	20	3	21	92	92	92	92	92	92	69.7	33.1	< 0.01

Monetary Variables

Description	Variable Name	Headquarters' Expected Length	Number of Long Responses	Percentile Lengths of Mail Long Responses							Mean	Standard Deviation	Truncation Rate (%)
				25th	50th	75th	90th	95th	99th	100th			
Self-employment income	SEM	6	351	7	7	7	7	7	8	10	7.1	0.3	0.16
Monthly insurance payment	INS	4	1,264	5	5	5	5	6	6	7	5.1	0.3	0.16
Social Security or Railroad Retirement income	SS	5	529	6	6	6	7	7	9	9	6.3	0.6	0.11
Supplemental security income	SSI	5	102	6	6	6	6	7	8	8	6.1	0.3	0.09
Yealy real estate tax	TAX	5	743	6	6	6	6	6	7	8	6.1	0.3	0.09
Yearly other fuel cost	FUL	4	226	5	5	5	5	6	6	6	5.1	0.2	0.06
Wages/ salary income	WAG	6	823	7	7	7	7	8	10	13	7.1	0.5	0.06
Interest Income	INT	6	260	7	7	7	7	7	8	9	7.1	0.3	0.05
Monthly condomium fee	CON	4	42	5	5	5	5	6	6	6	5.1	0.3	0.05
Yearly mobile home costs	MH	5	38	6	6	6	6	6	9	9	6.1	0.6	0.05
Yearly water and sewer costs	WAT	4	375	5	5	5	5	5	7	7	6.1	0.5	0.04
Monthly other mortgage payments	SM	5	77	6	6	6	6	6	9	9	5.1	0.3	0.04
Property value	VAL	7	346	8	8	9	9	9	10	11	8.3	0.5	0.04
Public assistance recipiency	PA	5	13	6	6	6	7	7	7	7	6.2	0.4	0.02
Monthly mortgage payment	MRG	5	96	6	6	6	6	6	7	7	6.0	0.2	0.02
Retirement income	RET	6	45	7	7	7	7	7	8	8	7.0	0.2	0.01
Monthly electricity cost	ELE	4	137	5	5	5	5	5	6	6	5.0	0.1	0.01
Monthly gas cost	GAS	4	84	5	5	5	5	5	5	5	5.0	0.0	0.01
Other income amount	OI	6	17	7	7	7	8	10	10	10	7.3	0.8	0.01
Monthly rent	RNT	5	21	6	6	6	7	10	10	10	6.4	1.1	0.01
Total income	TI	7	116	8	8	8	9	10	13	13	8.3	0.9	0.01

Note: The write-ins for field of degree, FODW, and health insurance, HINSW, are omitted here because there were no long responses for either item in the 2009 ACS sample.

By combining the results in Tables 2 and 3, we can see how expanding a variable's field length may decrease its long response rate. Table 4 shows what the 2009 unweighted truncation rates would have been if the Headquarters' expected length for each item was expanded to the 50th, 75th, and 90th percentile length of the item's long responses. For example, the American Indian or Alaska Native Race (RCW1) write-in, which has roughly a 2.9 percent long response rate, would have had about a 1.4 percent rate if its write-in field was expanded by 6 characters (its 50th percentile long response length).

Table 4. Potential Mail Mode Long Response Rates Based on Field Expansions, Unweighted

Source: 2009 American Community Survey Sample, Mail Keyed Data

Character Variables

Description	Variable Name	Headquarters' Expected Length	Character Length Expansion from Current Limit				Corresponding Truncation Rate (%)			
			Current	50th	75th	90th	Current	50th	75th	90th
American Indian or Alaska Native race	RCW1	30	0	6	11	16	2.89	1.44	0.72	0.29
Other Asian race or Pacific Islander	RCW2	30	0	6	11	19	1.11	0.56	0.28	0.11
Language spoken at home	LANW	20	0	2	6	11	1.11	0.56	0.28	0.11
Ancestry	ANCW	40	0	6	11	18	1.08	0.54	0.27	0.11
Hispanic origin	HISW	30	0	5	10	16	0.47	0.24	0.12	0.05
Migration foreign country	MGW1	30	0	9	21	24	0.39	0.19	0.10	0.04
Respondent first name	RFN	13	0	3	5	7	0.30	0.15	0.07	0.03
Second occupation	OCW2	60	0	9	20	38	0.27	0.13	0.07	0.03
Other race	RCW3	30	0	4	12	106	0.26	0.13	0.07	0.03
Place of work city	PWW2	20	0	4	9	17	0.24	0.12	0.06	0.02
Place of work county	PWW4	20	0	3	5	9	0.11	0.06	0.03	0.01
Migration city	MGW3	20	0	5	8	16	0.11	0.05	0.03	0.01
Place of work address	PWW1	53	0	7	14	25	0.09	0.05	0.02	0.01
Place of birth outside US	PBW3	30	0	3	7	15	0.08	0.04	0.02	0.01
First industry	INW2	60	0	8	15	26	0.05	0.03	0.01	0.01
Second industry	INW3	60	0	10	20	36	0.04	0.02	0.01	<0.01
Migration county	MGW4	20	0	4	5	8	0.04	0.02	0.01	<0.01
First occupation	OCW1	60	0	10	23	41	0.04	0.02	0.01	<0.01
Respondent last name	RLN	20	0	4	9	14	0.03	0.01	0.01	<0.01
Place of birth US	PBW2	20	0	5	10	18	0.02	0.01	0.01	<0.01
Migration address	MGW2	53	0	4	8	13	0.02	0.01	0.01	<0.01
Place of work state	PWW5	20	0	72	72	72	<0.01	<0.01	<0.01	<0.01
Migration state	MGW5	20	0	6	8	20	< 0.01	< 0.01	< 0.01	< 0.01

Monetary Variables

Description	Variable Name	Headquarters' Expected Length	Character Length Expansion from Current Limit				Corresponding Truncation Rate (%)			
			Current	50th	75th	90th	Current	50th	75th	90th
Self-employment income	SEM	6	0	1	1	1	0.16	0.08	0.04	0.02
Monthly insurance payment	INS	4	0	1	1	1	0.16	0.08	0.04	0.02
Social Security or Railroad Retirement income	SS	5	0	1	1	2	0.11	0.05	0.03	0.01
Supplemental security income	SSI	5	0	1	1	1	0.09	0.05	0.02	0.01
Yealy real estate tax	TAX	5	0	1	1	1	0.09	0.04	0.02	0.01
Yearly other fuel cost	FUL	4	0	1	1	1	0.06	0.03	0.02	0.01
Wages/ salary income	WAG	6	0	1	1	1	0.06	0.03	0.01	0.01
Interest Income	INT	6	0	1	1	1	0.05	0.03	0.01	0.01
Monthly condomium fee	CON	4	0	1	1	1	0.05	0.02	0.01	<0.01
Yearly mobile home costs	MH	5	0	1	1	1	0.05	0.02	0.01	<0.01
Yearly water and sewer costs	WAT	4	0	1	1	1	0.04	0.02	0.01	<0.01
Monthly other mortgage payments	SM	5	0	1	1	1	0.04	0.02	0.01	<0.01
Property value	VAL	7	0	1	2	2	0.04	0.02	0.01	<0.01
Public assistance recipiency	PA	5	0	1	1	2	0.02	0.01	<0.01	<0.01
Monthly mortgage payment	MRG	5	0	1	1	1	0.02	0.01	<0.01	<0.01
Retirement income	RET	6	0	1	1	1	0.01	0.01	<0.01	<0.01
Monthly electricity cost	ELE	4	0	1	1	1	0.01	0.01	<0.01	<0.01
Monthly gas cost	GAS	4	0	1	1	1	0.01	0.01	<0.01	<0.01
Other income amount	OI	6	0	1	1	2	0.01	0.01	<0.01	<0.01
Monthly rent	RNT	5	0	1	1	2	0.01	<0.01	<0.01	<0.01
Total income	TI	7	0	1	1	2	0.01	< 0.01	< 0.01	< 0.01

Note: The write-ins for field of degree, FODW, and health insurance, HINSW, are omitted here because there were no long responses for either item in the 2009 ACS sample.

It appears that a small increase in the maximum character length (0 to 6 characters), for most items in 2009, would have eliminated a majority of truncation. However, knowing whether the extra characters kept would make a difference in terms of data quality requires further research.

4. How often do open-ended responses in the automated modes equal their maximum character length in data collection?

As discussed in the introduction section, automated mode responses are not subject to truncation because interviewers collect data using the CATI and CAPI instruments. Interviewers know they have reached an item's maximum size when the instrument does not let them type any more into a field. Here we assess how often data collected by CATI and CAPI interviewers are close to or meet their item specific maximum field lengths.

Table 5 summarizes the distribution of response lengths for each item using the percentiles, mean, and standard deviation of their long response lengths. The 50th percentile length equals the median length and the 100th percentile length equals the maximum length. When the 100th percentile value equals the Headquarters' expectation we know that interviewers reached the maximum input length allowed in the instrument and may have had to truncate or abbreviate a response.

Eight character items had at least one response as long as the maximum length. However, most of these items had 90 percent of their responses shorter than their maximum lengths by 8 characters or more. The two exceptions are the Reference First Name (RFN) and Migration address (MGW2) item. These questions had 90 percent of their long responses shorter than their maximum lengths by only 5 characters and 3 characters, respectively. Please note that RFN is a proxy for all the first name fields on the ACS questionnaire. We discuss this in the limitations section.

For the monetary items, nearly all items have at least one response equal to their maximum field length. Three items actually exceed their maximum length because ACS processing allows negative monetary values an additional character to account for the negative ("-") sign. However, these are not necessarily major concerns because, for example, a four-digit field can hold values ranging from $1 to $9,999 dollars. If very few responses are near $9,999 then having only four digits is not a problem. This is the case for most of the monetary items.

The Social Security and Railroad Retirement (SS) and monthly insurance payment (INS) were the only two monetary items that had responses in their respective 75th percentile equal their maximum length in the 75th percentile. This means 25 percent of SS and INS responses used the full field. The SS write-in has a five character limit and INS has a four character limit; however, only 0.1 percent and 0.6 percent of the responses for these variables, respectively, had to be top coded.[2]

[2] Monetary amounts are top coded when the number of digits in the amount exceeds the maximum number of allowable characters. For example, a response of $10,000 to an item with a four character maximum length would be top coded as $9,999.

Table 5. Automated Modes' Response Length Distribution, Unweighted

Source: 2009 American Community Survey Sample, Data Capture File

Character Variables

Description	Variable	Total Responses	Mean	Standard Deviation	Percentile Length 25th	50th	75th	90th	95th	99th	100th	Headquarters' Expectation
Respondent first name	RFN	2,409,760	6.2	1.6	5	6	7	8	9	11	13	13
Respondent last name	RLN	2,409,760	7.0	2.1	6	7	8	9	11	14	20	20
Place of birth US	PBW2	1,603,297	8.4	3.5	7	8	10	13	13	17	20	53
Place of birth outside US	PBW3	1,603,297	8.4	3.3	6	7	10	12	17	18	20	30
Ancestry	ANCW	1,495,751	12.6	4.5	8	13	17	18	18	20	20	40
First industry write-in	INW2	933,974	15.0	3.7	13	17	18	18	19	19	20	60
Second industry write-in	INW3	933,974	13.3	4.1	10	13	17	18	19	19	20	60
First occupation write-in	OCW1	933,974	13.3	4.2	10	14	17	18	19	19	20	60
Second occupation write-in	OCW2	933,974	15.2	3.5	13	17	18	18	19	19	20	60
Place of work address	PWW1	722,401	13.5	3.2	11	14	16	17	18	18	20	53
Place of work city	PWW2	722,401	9.1	2.5	7	9	11	12	13	16	20	20
Place of work county	PWW4	722,401	7.9	2.4	6	8	9	11	12	15	20	20
Place of work state	PWW5	722,401	7.1	3.7	2	8	10	12	13	13	19	20
Language spoken at home	LANW	310,619	7.8	2.5	7	7	7	10	14	18	20	20
Field of Degree	FODW	292,483	15.0	4.2	11	17	18	19	19	20	20	75
Migration foreign country	MGW1	142,857	8.5	3.8	6	7	11	15	17	19	19	30
Migration address	MGW2	142,857	13.7	2.8	12	14	16	17	17	18	20	20
Migration city	MGW3	142,857	9.0	2.5	7	9	10	12	13	15	20	20
Migration county	MGW4	142,857	7.8	2.4	6	8	9	10	12	14	20	20
Migration state	MGW5	142,857	7.2	3.5	4	8	10	12	13	13	17	20
American Indian or Alaska Native Race	RCW1	131,213	10.7	4.3	8	9	15	18	18	19	20	30
Other Asian Race or Pacific Islander	RCW2	131,213	9.7	3.7	7	8	12	16	17	19	20	30
Other Race	RCW3*	131,213	9.4	3.5	7	9	11	15	17	18	20	30
Hispanic origin write-in	HISW	62,369	10.3	3.2	8	10	11	16	18	19	20	30
Other health Insurance	HINSW	52,471	13.6	4.5	10	15	17	18	18	19	20	30

Monetary Variables

Description	Variable	Total Responses	Mean	Standard Deviation	Percentile Length 25th	50th	75th	90th	95th	99th	100th	Headquarters' Expectation
Interest Income	INT	4,958,571	1.9	1.5	1	1	3	5	5	6	7	6**
Other income amount	OI	4,958,571	1.5	1.2	1	1	1	4	5	5	6	6
Publis assistance recipiency	PA	4,958,571	1.1	0.5	1	1	1	1	1	4	5	5
Retirement Income	RET	4,958,571	2.0	1.6	1	1	4	5	5	5	6	6
Self-employment income	SEM	4,958,571	1.7	1.5	1	1	1	5	5	6	7	6**
Social Security or Railroad Retirement inc	SS	4,958,571	2.5	1.8	1	1	5	5	5	5	5	5
Supplemental security income	SSI	4,958,571	1.3	0.9	1	1	1	1	4	5	5	5
Total income	TI	4,958,571	4.4	1.4	4	5	5	5	6	6	8	7**
Wages/ salary income	WAG	4,958,571	4.0	1.7	3	5	5	5	6	6	6	6
Monthly condominium fee	CON	2,409,760	1.2	0.6	1	1	1	2	3	3	4	4
Monthly electricity cost	ELE	2,409,760	2.4	0.7	2	3	3	3	3	3	4	4
Yearly fuel cost	FUL	2,409,760	1.7	1.2	1	1	3	4	4	4	4	4
Monthly gas cost	GAS	2,409,760	1.9	0.8	1	2	3	3	3	3	4	4
Monthly insurance payment	INS	2,409,760	2.7	1.1	1	3	4	4	4	4	4	4
Yearly mobile home costs	MH	2,409,760	1.3	0.8	1	1	1	3	4	4	5	5
Monthly mortgage payment	MRG	2,409,760	2.6	1.3	1	3	4	4	4	4	5	5
Monthly rent	RNT	2,409,760	2.0	1.2	1	1	3	4	4	4	5	5
Monthly other mortgage payments	SM	2,409,760	1.5	0.9	1	1	1	3	3	4	5	4
Yearly real estate tax	TAX	2,409,760	3.2	1.2	3	4	4	4	4	5	5	5
Property value	VAL	2,409,760	5.2	1.7	5	6	6	6	6	7	7	7
Yearly water and sewer cost	WAT	2,409,760	2.5	1.0	1	3	3	4	4	4	4	5

5. Would expanding the maximum character lengths in data collection or subsequent data processing for certain survey items capture more meaningful data?

For the mail mode, ACS may consider the four variables having truncation rates above 1 percent for the possibility of field expansion. These variables the ancestry write-in (ANCW), the language spoken at home write-in (LANW), the American Indian or Alaska Native Race write-in (RCW1), and the Other Asian Race or Pacific Islander Race write-in (RCW2). Because the responses to each of these items are subject to automated and clerical coding, ACS would need to consider if increasing the length of strings sent to coding would change the values of and/or number of codes assigned to each response.

In all coding operations, responses are assigned a certain number of "codes". In the ancestry coding operation, for example, coders issue a maximum of two codes per response. If a respondent provides more than two ancestral origins, then the coder is unable to record the additional ancestries. Thus, increasing the length of the string passed to the coding operations may not change the resulting data for an item unless the ACS increased the number of recordable codes.

In the automated modes, most responses are not truncated (there was one anomaly discussed in results for research question #1). Interviewers are aware of the maximum input lengths because they record respondent answers directly into the CATI and CAPI instruments. It is possible that interviewers abbreviate longer responses in order to fit them into the answer field. We have no measure of how often this actually happens.

To determine if the ACS would collect more meaningful data by expanding the maximum character lengths, the survey should first inquire with coders how often they are unable to code a string due to abbreviations. If it appears to be a legitimate factor, the ACS should investigate whether allowing longer write-ins changes the values of or number of codes assigned to long responses.

Conclusions

Most mail mode items have low rates of truncation; only four had rates above 1.0 percent: the ancestry write-in (ANCW), the language spoken at home item (LANW), the American Indian or Alaska Native Race write-in (RCW1), and the Other Asian Race or Pacific Islander Race write-in (RCW2). The volume of long responses is also low for most items (less than 4,300 annually).

The median long response length for LANW is 2 characters longer than the expected length, and for RCW1, RCW2, and ANCW their median long response lengths are about 6 characters longer than the limit. This suggests that mail processing may truncate the end of a word for LANW and, for RCW1, RCW2, and ANCW it may truncate a whole word or the end of a word. However, as discussed in research question #5, this is inconclusive evidence that expanding the maximum character lengths would capture data that are more meaningful. The ACS also would need to consider the number of codes that can be assigned.

In the automated modes, unlike in the mail mode, truncation is less of a concern because the interviewers know the maximum field sizes permitted by the data collection instruments and may

adjust long responses accordingly. However, in conducting this research, we found one anomaly. The Other Pacific Islander Race responses in the automated instruments are truncated if the string for the Other Race item contains data and the length of the concatenated responses for the two items are more than 30 characters (see Appendix A for the full explanation). This only happened 9 times in the course of three years. So, although it is not an immediate issue, we should fix this issue.

The majority of character responses in the automated modes are shorter than their maximum lengths by 8 characters or more. The Respondent First Name (RFN) and Migration address (MGW2) items, however, have responses shorter than their maximum lengths by only 5 characters and 3 characters, respectively. Please note that the RFN variable is a proxy for all the first name fields on the ACS questionnaire (discussed in the limitations section). Additionally, the monetary items all seem to fit into their allowable lengths. Although they each have at least one response equal to the maximum field length, their values are far below their maximum allowable value.

As is the case with the mail mode, additional research into how coded responses change when a longer string is used would be necessary to state whether expanding the item field lengths would capture data that are more meaningful. A separate study analyzing how coding outcomes (number of codes or values of codes) change by using longer maximum lengths would help answer this question.

Because the ACS plans to redesign its Headquarters' system as a result of the Bureau's ongoing efforts related to adaptive design, we recommend that the redesign look into the feasibility of expanding the Headquarters' expected lengths.

Appendix A

The Other Pacific Islander race, RCW4, which is a field by itself only in the automated instrument, poses a potential data collection problem. Before an automated mode response to this item is sent to coding it is concatenated to the end of the Other Race open-response item, RCW3. The resulting string is truncated to the Headquarters' expected length for RCW3, which is currently 30 characters. Thus, it is possible that information stored in RCW4 would never be sent to coding and be lost from this point forward.

To see how often data are actually lost, we tallied how many conjoined RCW3 + RCW4 responses were longer than 30 characters using the data capture file (DCF). In essence, we counted the number of long responses to RCW3 + RCW4. We found that throughout the 2008, 2009, and 2010 ACS samples, there were only 83 long responses to RCW3 + RCW4. And, all 83 had a concatenated length of 31 characters. See the Table A below:

Table A. Frequency and Lengths of RCW3 + RCW4 Long Responses

ACS Sample Year	Number of RCW3 + RCW4 Truncations	Length of All RCW3 + RCW4 Responses	Number of RCW3 + RCW4 Truncations Sent to Coding
2008	18	31	1
2009	39	31	5
2010	26	31	3
Total	83	31	9

These 83 long responses resulted from two response patterns to RCW3 and RCW4. First, there were 79 cases that answered "don't know" or "refused" to RCW3, which took up 30 characters-- 29 blank spaces with the 30th character as a "D" or "R", and had RCW4 blank or "D", which took up 1 character. Here's an example with underscores used to represent individual spaces:

RCW3	+	RCW4	=	RCW3 + RCW4
"_____R"		"D"		"_____RD"

These cases had a concatenated length of 31, but were not sent to coding. The end result for these cases on the DCF showed the new RCW3 response as a blank, which is what we'd want since the respondent indicated a "don't know" or "refused" response to both items.

Second, there were 4 cases out of the 83 that answered the original RCW3 or RCW4 with a specific 30 character response and a blank response (a one character space, ie. " ") to the other item. All of these cases were sent to coding and the resulting new RCW3 on the DCF matched the 30 character response of the original RCW3 or RCW4.

So, in all, it does not appear that the ACS has lost data in the past few years; however, this may be an undesired method of processing and could be addressed in future modifications to the ACS data processing systems.