Utilizing Paradata to Examine and Improve the Web Data Collection Process in Agricultural Census and Survey Programs

Robyn Sirkis[1]
United States Department of Agriculture-National Agricultural Statistics Service,
1400 Independence Ave., SW, Washington, DC 20250

Abstract
In today's world, statisticians must consider and gain insights from multiple, and sometimes complex, data sources. One such source is paradata which are metrics related to the process of collecting survey data. In web surveys, paradata can be classified into several types, including device (e.g., device type, browser used), and questionnaire navigation (e.g., changing answers, error triggers, breakoffs). With an increasing focus on efficiencies in data collection, the USDA's National Agricultural Statistics Service (NASS) began collecting and analyzing paradata for its web data collections. This research was conducted to learn more about the web collection process at the respondent level and make informed improvements to the online questionnaire. The analysis includes a comprehensive evaluation across mobile and non-mobile device types. This paper focuses on the methodology and results using web paradata from the 2017 Census of Agriculture and June 2018 Agricultural Survey. Challenges encountered analyzing the paradata and recommendations based on our findings will also be discussed, illustrating the valuable role paradata can play in helping survey organizations.

Key Words: paradata, data collection, web instruments, questionnaire design, mobile, survey methods

1. Introduction

Many survey organizations are turning their attention to collecting survey data using the web. With rising data collection costs, utilizing the web is becoming increasingly popular. Collecting data using web instruments does not require interviewers, printing, or keying. Web surveys have many unique advantages, such as a variety of interactive features including help buttons and visually stimulating colors and graphics. Programmed skip patterns and screening questions grant the respondent the opportunity to focus on questions that specifically apply to them, therefore reducing burden. Warning screens can be programmed that allow for instantaneous feedback for unusual or missing responses that allow the respondent to modify their answers. Increasing your focus to using web forms does not come without some considerations. The use of the web is dependent on the willingness of the respondent to complete the survey (Couper, 2000), internet accessibility with a rural population, and the data quality. Analysis of paradata can help evaluate the impact of the unique features of web instruments.

[1] The findings and conclusions in this publication are those of the author and should not be construed to represent any official USDA or U.S. Government determination or policy.

2. Background

2.1 Paradata

Paradata are data about the process of answering the survey itself (Couper, 2000). Web surveys are self-administered, meaning the paradata are generated by respondents and their interaction with survey instruments (Callegaro, 2013). Paradata are not consciously provided by the respondents, but are collected from the respondent's computer or device (Kaczmirek, 2008). Callegaro (2013) categorized paradata into two classes, device type paradata and questionnaire navigation paradata. Device type paradata includes a variety of information such as the browser used, mobile device used (e.g., iPad), and browser window size. This type of data are provided at the session level with the possibility that respondents may complete the survey in one or multiple sessions. Questionnaire navigation paradata refers to data about the process of filling out the questionnaire. Examples include the respondent changing answers, appearance of warning messages, clicks on non-questions (e.g., save and help buttons), and the last question the respondent answered if they exited the web form early. This type of data are collected at the question or page level. This paper discusses paradata collected in two NASS programs, the Census of Agriculture and the June Agricultural Survey.

If paradata is shown to be useful to survey organizations, they may also want to use information collected to increase the amount of data collected by web in the future. Much descriptive information about agricultural operations and their operators is collected in the Census of Agriculture. This data can be used to build models to predict the types of operations most likely to respond via web. This paper will also discuss work to develop these models.

2.2 Census of Agriculture 2017

The Census of Agriculture (COA) is a complete count taken every five years of U.S. farms and ranches and the people who operate them. The questionnaire has thirty-four sections on topics such as land use and ownership, agricultural production, operator characteristics, demographics, production practices, income, and expenditures. The COA provides the only source of uniform, comprehensive, and impartial agriculture data for every county in the nation. The 2017 COA data collection period was from November 2017 to August 2018 (COA, 2017). The respondent could complete the form by mail or web. Follow-up data collection methods included Computer Assisted Telephone Interviewing (CATI) and Computer Assisted Personal Interviewing (CAPI). The total census mail list count was approximately 3 million records, with a final count of approximately 2 million farms. Approximately 70 percent of respondents completed the 2017 COA by mail, 23 percent by web, and 2 percent by the interviewer administered modes. The estimated time required to complete the survey is on average 50 minutes.

2.3 June 2018 Agricultural Survey

The June 2018 Agricultural Survey provides estimates of crop acreage, yields, and production and quantities of grain and oilseeds stored on farms. The Agricultural Surveys (one of them is in June) are conducted in all states quarterly. Data are used by commodity markets, educational institutions, state, and federal agencies, farm and ranch operators, and others for market assessment, planning, decision making and ongoing research. Total sample sizes range from about 65,000 in September to about 81,000 in June partially because the number of crops of interest varies between quarters. Modes of data collection include mail, web, telephone, CATI, and personal interviewing (Crops/Stocks, 2018). The estimated time required to complete the survey is on average 25 minutes.

3. Methodology

3.1 Device Type Paradata

Survey organizations could greatly benefit from analyzing device type paradata from their web instruments. The percent of respondents by device (e.g., phone), browser, and mobile device (e.g., iPhone, iPad) was calculated for the 2017 COA and June 2018 Agricultural Survey. One advantage of analyzing this information is that questionnaire designers can focus on the top used device and browsers when trying to optimize the capability and design of the web form. The cooperation rate by device was also calculated for both surveys, which is the percent of respondents that submitted the web form divided by the percent of respondents that accessed the web form. The cooperation rate provides an indication if the respondents are exiting the web form early. However, item nonresponse must be taken into account before making an overall statement regarding data quality and ease of use. The cooperation rate was also calculated by device to determine if there might be some issues, particularly with phones due to smaller screen sizes. The percent of respondents that switched devices was also calculated to determine if respondents were having difficulty completing the web form on their chosen device.

3.2 Questionnaire Navigation Paradata

Questionnaire navigation paradata refers to the process of filling out the questionnaire. This can include changing answers, clicks on non-questions (e.g., help button), and the last question answered if the respondent exited the form early (Callegro, 2003). This provides valuable information since it can indicate whether problems exist at the question level. The percent of break offs by section, percent of changed answers by question, and the number of respondents accessing the help button was calculated for the 2017 COA. The percent of changed answers by question and the number of times respondents clicked on help was calculated for the June 2018 Agricultural Survey. There were not many break offs or respondents who clicked the help button in the June 2018 Agricultural Survey. It is necessary to examine break offs since it is possible that respondents can exit early due to sensitivity or confusion of the questions. A closer look at questions where respondents changed their answers the most is also warranted to determine if there are better methods to administering these items. Respondents can change their answers for a number of reasons. They can change their answers due to misunderstanding the content of the question or attempting to reconcile between answers in the same section or across sections if warning messages exist. Respondents can change their answers to screening questions if they want to avoid answering additional questions. Cognitive and usability testing are methods that can be utilized to understand how to modify or remove problematic questions.

3.3 Effectiveness of Warning Screens

Warning screens can be beneficial to the overall data quality of a web survey. They may minimize item nonresponse if the respondent forgot to answer a question tied to a warning screen. They also provide an opportunity for respondents to modify incorrect responses. There are different types of warning messages that can appear in a web form. Warning messages can be used to reconcile answers in the same section or across sections. For example, in the 2017 COA, one of the warning messages is “the total land use acres do not equal the total acres operated reported in the previous section on acreage”. Another example is “the total acres operated by county exceeds the total acres operated for this operation”. Other warning messages have the purpose of ensuring the respondent fills in an answer such as “please enter a response”. Some warning messages are placed in web forms to correct grammatical errors such as “please enter a valid email address” or “please enter a 4-digit year prior to 2017”. The “@” symbol can be accidently omitted from a

respondent's email address response. Respondents also may inadvertently enter a two-digit or three-digit year instead of the requested 4-digit year.

There were 10 warning messages in the COA web form. To determine which messages the respondents triggered most often, the percent the warning message was triggered out of the total warnings was calculated. Then the final response was analyzed to determine whether the respondent potentially changed their answer due to the warning message. This would be an indication of the effectiveness of the messages. The warning messages information was on the COA web paradata files, but not available for the June 2018 Agricultural Survey.

3.4 Item Nonresponse

Item nonresponse rates were compared between web and mail and then web and CATI using the 2017 COA. The number of web respondents was much larger for the 2017 COA than the June 2018 Agricultural Survey. Therefore, the item nonresponse and modeling analysis was only conducted for the 2017 COA. The chi-square test was used to determine whether significant differences exist between web and each of the other modes. The null hypothesis is that the proportions are equal. Twenty-six variables were selected on a variety of topics such as demographics and the commodity screening questions as shown in Table 1. The commodity screening questions were chosen since the format is vastly different between the web and mail modes. Figure 1 shows a portion of the web screen displaying the crop commodity screening questions. Six crop commodity screening questions appear on one web screen. If the respondent selects "yes", questions pertaining to that commodity will display on later screens. Figure 2 shows a few of the questions for two sections on the mail form. The screening question is at the top of each section. The theory was that more respondents would fill in the questions on the crop and livestock commodity screens in the web form since they are grouped together.

Table 1: Item Nonresponse (26 variables)

• Age	• Household size
• Sex	• Year started operating any farm
• Ethnicity	• Year started operating this farm
• Race	• Access to internet screener
• Days worked off farm	• Six crop commodity screeners
• Retired	• Eight livestock commodity screeners
• Principal occupation	• Total number of commodities

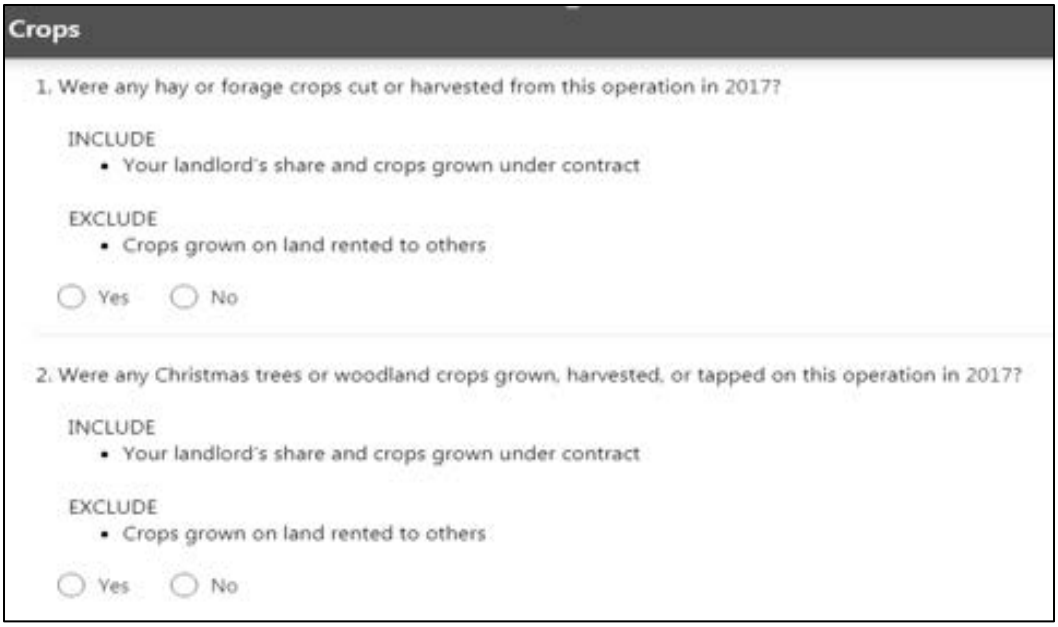

Figure 1: Web Form Yes/No Screening Questions

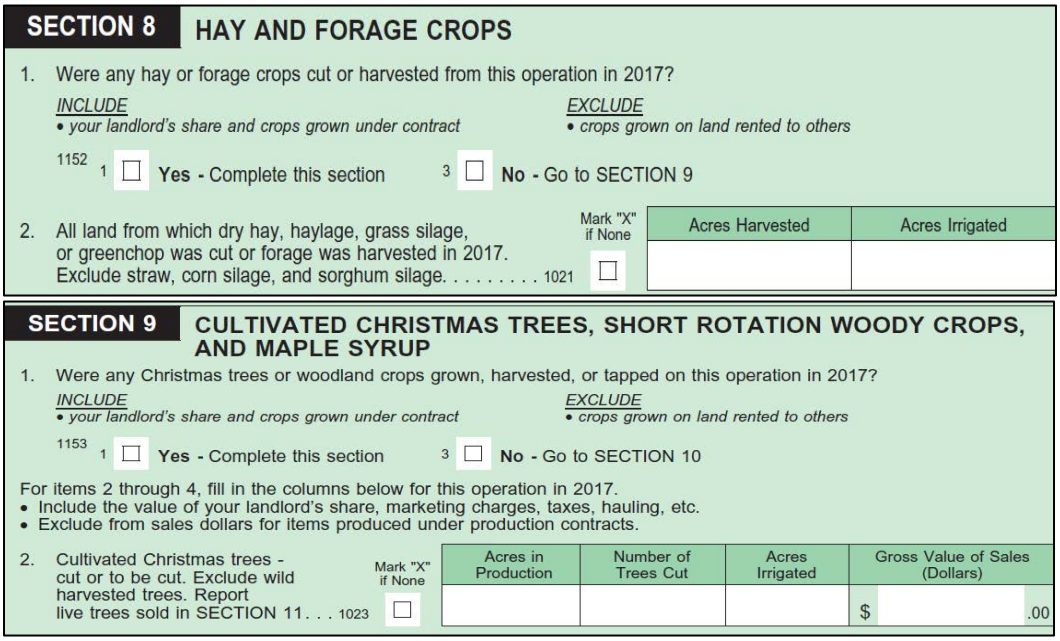

Figure 2: Mail Form Yes/No Crop Screening Questions

3.5 Modeling

The COA collects a number of variables about the respondents' agricultural operation and its operators. This information can be used to describe and identify differences between respondents by mode. The majority of respondents completed the 2017 COA by mail. A future goal is to determine characteristics related to web respondents to target those who completed the COA by mail or the interviewer assisted data collection modes to encourage web response. Eleven variables, as shown in Table 2, in the COA were selected that were expected to show significant differences between web and mail. The focus was on mail since it is also self-administered and 70 percent of the respondents completed the form

using this data collection method. The percentage of responses was calculated for each category for each variable to determine the extent of the differences between the web and mail modes.

Table 2: COA Characteristics Analyzed

• Age	• Principal occupation
• Sex	• Household size
• Ethnicity	• Access to internet
• Race	• Total number of commodities
• Days worked off farm	• Burden measure (number of sections)
• Retired	

Bootstrap forest models were used to identify additional potential characteristics related to the mode of response. The response variable for one of the models was web versus mail. The second model was web versus all modes (mail, CATI, CAPI). Logistic regression models were then developed to identify additional informative metrics related to those predictor variables.

The Bootstrap forest method was used since it generates many decision trees and averages the predicted values to get the final predictions. The minimum and maximum splits per tree for the initial model were set at 10 and 2000, respectively. Eight predictors were set to be sampled at each split. At least 1,104 observations were needed for the web versus mail mode model and 1,129 for the web versus all the other modes model at each tree node for it to be further split. Allowance of early stopping was also set for the bootstrap forest, which was employed in the selected model. Figures 3 and 4 show the bootstrap forest specifications. The validation and training data sets were used in the models. The validation set evaluates how well the model fits. It also is a measure of how well the model will fit on new observations. The training set estimates the model parameters and how well the model partitions the data. The Receiver Operating Characteristic (ROC) curve was analyzed as an indicator of the goodness of fit for the model. A value of 1 under the curve indicates a perfect fit and a value near 0.5 indicates that the model cannot discriminate among groups. The column contributions show which variables most influence the response variable, including their contribution to the fit. Table 3 shows the thirty-two variables selected for the initial models covering a variety of topics such as demographics, type of internet, and estimated value of all products produced as reported in the COA form. Additional variables that are not from the COA were included such as the response history and internet speed.

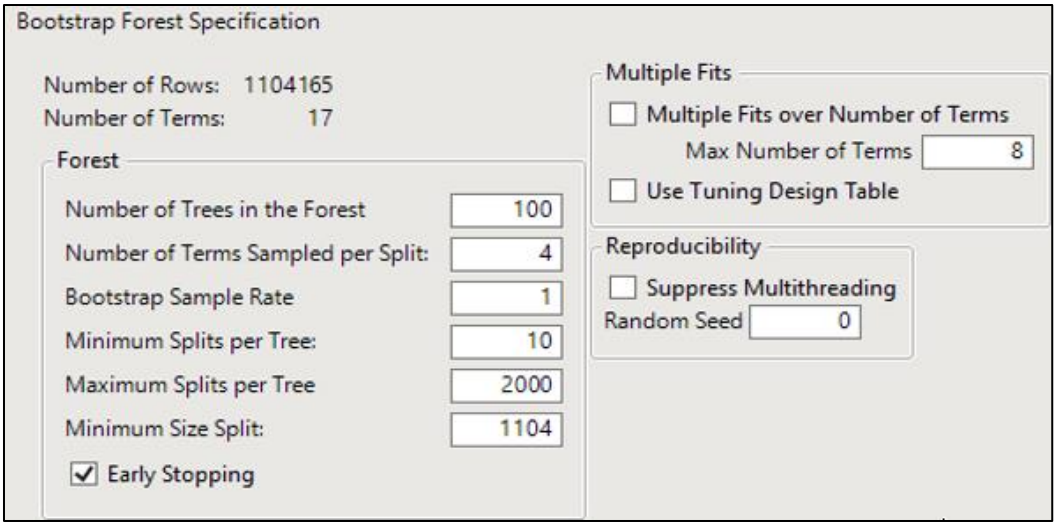

Figure 3: Web versus Mail Bootstrap Forest Specifications Final Model

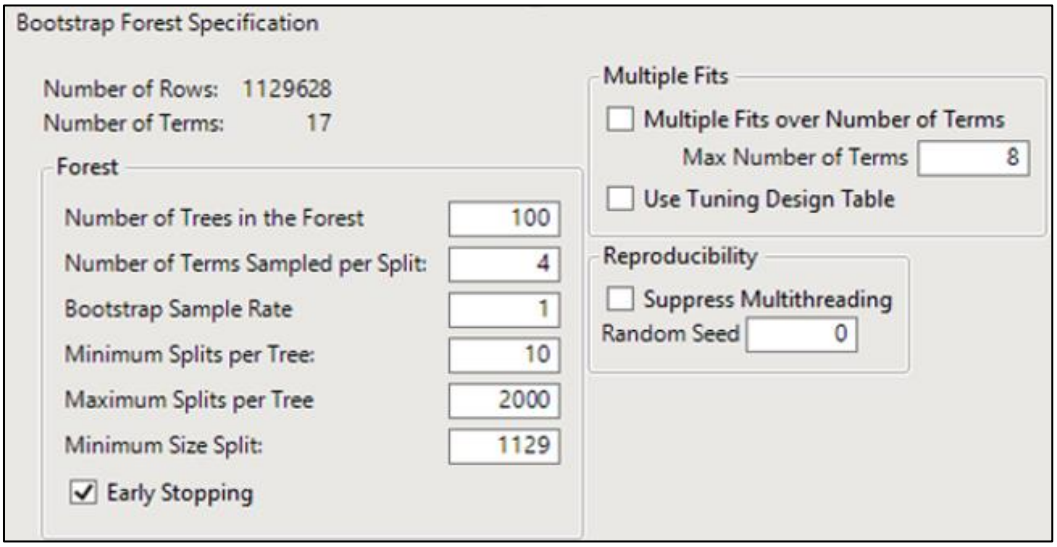

Figure 4: Web versus Mail/CATI/CAPI Bootstrap Forest Specifications Final Model

Table 3: Variables Chosen for the Model (32 variables)

• Age	• Region
• Sex	• Total acres
• Ethnicity	• Farm Type
• Race	• Hybrid of farm type/NAICS code
• Days worked off farm	• Estimated value products produced
• Retired	• Gross cash farm income
• Principal Occupation	• Number crop/livestock commodities
• Household size	• Burden measure (number of sections)
• Access to the internet	• Number of completed surveys 2 years
• Internet type: cable	• Number of completed surveys 3 years
• Internet type: dsl	• Number of completed surveys 4 years
• Internet type: mobile	• Number of completed surveys 5 years
• Internet type: satellite	• Total number of surveys 2 years
• Residential high-speed internet connections	• Total number of surveys 3 years
• Non-residential high-speed internet connections	• Total number of surveys 4 years
• Total high-speed internet connections	• Total number of surveys 5 years

Respondent level logistic regression models were created using the highest contributors from the bootstrap forest method. In addition, only main effects were considered in the potential statistical models. Issues relating to sample size, missing values, and collinearity among predictor variables were taken into account when choosing the covariates. Automatic selection methods such as forward selection were used to aid in choosing a parsimonious model. Potential models were evaluated based on model fit statistics such as the R-squared value, which refers to the fraction of variance explained by the model. The odds ratios were examined for those covariates left in the statistical model. The odds ratio is the probability of an occurrence of an event to that of non-occurrence. It accesses the strength of association and the potential impact of confounding variables.

There are characteristics that might be different between web and mail respondents that were not included in the statistical models due to the lack of availability of data. This includes preference of mode, technical ability using the web, and possible confidentiality concerns of submitting their personal information over the internet.

4. Results

4.1 Device Type Paradata

The majority of respondents used the desktop/laptop for both web surveys. Table 4 shows the percentage of respondents using each device type. The percent of respondents who used the desktop or laptop was 83.6 for the 2017 COA and 91.6 for the June 2018 Agricultural Survey. The percent of respondents who switched devices was 5.3 for those completing the 2017 COA web form in more than one session. Approximately 56 percent of those respondents switched from a mobile device to the desktop/laptop. Only six respondents switched devices in the June 2018 Agricultural Survey.

Table 4: Device Types

	2017 COA			June 2018 Agricultural Survey	
Device	**Number**	**Percent**		**Number**	**Percent**
Desktop/Laptop	389,356	83.6		1,176	91.6
Tablet	46,387	10.0		79	6.2
Phone	28,370	6.1		28	2.2
E-Book Reader	1,729	0.4		1	0.1
Unknown/Other	39	0.0		0	0.0

The cooperation rates were high for both web surveys. The cooperation rate was 96.3 percent for the 2017 COA and 95.1 percent for the June 2018 Agricultural Survey. The break off rate was less than 5 percent for both surveys. Generally, once respondents accessed the web form they submitted it. This might indicate that the web form was not difficult to use once they started it, however item nonresponse must be considered before making an overall statement regarding the usability. The cooperation rate was the lowest for the phone device, which could be due to the small screen size. In addition, the 2017 COA has grid sections for the personal characteristics and commodity sections which might not fit perfectly on a phone screen. Table 5 shows the cooperation rates by device. The other/unknowns consist of respondents where it was difficult to identify their device or device types that were not used by many respondents (such as a PlayStation or TV). Related to the cooperation rate, approximately 80 percent of respondents completed the web form in one session for the 2017 COA and 97 percent for the June 2018 Agricultural Survey. The June 2018 Agricultural Survey is shorter on average with an expected completion time of 25 minutes compared to the 2017 COA with an average completion time of 50 minutes. The counts for the number who accessed the form will differ slightly from Table 4 since there were respondent records on the device type paradata files that were not on the navigational based paradata files.

Table 5: Cooperation Rates

	2017 COA			June 2018 Agricultural Survey	
Device	**Number Accessed Form**	**Cooperation Rate**		**Number Accessed Form**	**Cooperation Rate**
Desktop/Laptop	389,352	96.5		1,176	95.0
Tablet	46,387	96.1		79	93.7
Phone	28,369	94.0		28	92.9
E-Book Reader	1,729	96.4		1	100.0
Other/Unknown	846	99.1		50	100.0
Total	466,683	96.3		1,334	95.1

The top mobile device type (i.e. phone or tablet) that respondents used to complete the 2017 COA web form was the iPad with 37.0 percent followed by a non-iPad tablet with 23.7 percent. However, the top mobile device used to complete the June 2018 Agricultural Survey was non-iPad with 42.6 percent followed by iPad with 32.4 percent. As expected the majority of respondents used the tablet in the landscape orientation and the phone in the portrait orientation. Over 70 percent of respondents across both surveys used the landscape orientation when completing the web form using the tablet. More than 85 percent

of respondents across both surveys used the portrait orientation when completing the web form using the phone. Table 6 shows the percentage of respondents using each mobile device type. Table 7 provides the percentages of respondents by mobile device type and the orientation that they held the device.

Table 6: Mobile Devices

Type	2017 COA Percent	June 2018 Agricultural Survey Percent
iPad	37.0	32.4
Other Tablet	23.7	42.6
iPhone	20.5	13.9
Android Phone	12.5	9.3
Other	6.3	1.9

Table 7: Mobile Orientation

	2017 COA			June 2018 Agricultural Survey	
Orientation	**Phone**	**Tablet**		**Phone**	**Tablet**
Landscape	9.1	75.5		10.7	78.5
Portrait	90.9	24.4		89.3	21.5

More than half of the respondents used Chrome and Internet Explorer to complete the web form for both surveys. This is important since questionnaire designers can place their initial focus on the top browsers used by respondents since testing time might be limited. Table 8 shows the percentage of respondents that used each browser type.

Table 8: Browser Types

Browser	2017 COA Percent	June 2018 Agricultural Survey Percent
Chrome	40.9	43.8
Internet Explorer	17.9	17.8
Safari	16.3	9.8
Edge	13.4	17.4
Firefox	10.2	10.7
Other	1.2	0.6

4.2 Navigational Based Paradata

As stated previously, less than five percent of respondents for both surveys exited the survey without submitting the form. However, it is informative to know where in the web form respondents broke off to assess which questions may be causing confusion, frustration, or irritation. The top section where respondents broke off in the 2017 COA was the production expenses section with 23.5 percent, which is near the end of the questionnaire. The section asks for exact dollar amounts across a wide variety of expenses, which might be a sensitive, difficult, or burdensome topic for some respondents. This was followed by the Acreage in 2017 section (the first content related section) with 11.3 percent and the Out of Business Screeners section (which appears before any survey content to determine survey eligibility) with 10.0 percent, both at the beginning of the web form.

The questions where more respondents changed their answers in the 2017 COA were about acreage and those that required reconciling within the section and across sections and were tied to warning messages. For example, as shown in Table 10, the majority of the questions in the land use section had to be reconciled with the previous section on acres operated. The question where most respondents changed their answer in relation to those that answered was an aquaculture "other write-in" question. This is likely a function of the questionnaire content, because the instructions for what to include and exclude in this section may not be clear. The top questions changed in the June 2018 Agriculture survey as shown in Table 9 were related to acreage.

Table 9: Percent Changed Answers (Agricultural Survey June 2018)

Section Name	Question Description	Percent of Positive Responses Changed
Acres Operated	Acres Used On A Fee Per Head Or AUM Basis	10.2
Conclusion	Day-to-day decisions for another farm or ranch	10.1
Acres Operated	Land Owned	9.8
Acres Operated	Land Rented From Others	9.6

Table 10: Percent Changed Answers (Census of Agriculture 2017)

Section Name	Question Description	Percent of Positive Responses Changed
Aquaculture	Aquaculture specify	36.0
Land Use	Permanent Pasture and Rangeland, Acres	20.6
Cattle and Calves	Beef Cow--Inventory	18.8
Land Use	Cropland Harvested, Acres	18.4
Land Use	All Other Land, Acres	18.2
Vegetables	Land Used for Vegetables Harvested, Tenth-Acres	14.4
Acreage	Owned Land Rented to Others, Acres	14.4
Land Use	Woodland Not Pastured, Acres	13.3
Land Use	Woodland Pastured, Acres	10.7

Another indication there might be a problematic question or section is the number of respondents that clicked the help button. This analysis was conducted using the 2017 COA. Approximately 5 percent of respondents selected the help button one or more times. Approximately 40 percent of the respondents that selected help did so at the beginning of the session during the Out-of-Business Screener or Acreage in 2017 sections.

4.3 Warning Screens

There are 10 warning messages in the 2017 COA web form. The two warning messages triggered the most were "the total land use acres do not equal the total acres operated" and "please enter a valid e-mail address". The number of respondents who triggered at least one warning message was 46 percent. Out of all the warning messages, approximately 42 percent of the respondents triggered the warning "the total land use acres do not equal the total acres operated". Respondents have to make sure their total acres operated in section one equals the total land use acres in section two. It is not as simple as having two numbers match since the totals are automatically calculated from the sum of several questions. The

warning message "please enter a valid email address" was triggered by 31 percent of the respondents who triggered at least one warning. Approximately 38 percent of the respondents who triggered at least one warning left the answer field blank. In terms of the effectiveness of the warning messages, for 9 out of 10 messages at least 74 percent of respondents fixed their responses. Table 11 shows the percent of respondents triggering each warning message compared to all warning messages and the percent of respondents who modified their responses. From these results, it appears that warning messages are improving data quality by prompting respondents to correct their answers.

Table 11: Effectiveness of Warning Messages

	Respondents		
Warning Message	**Total[2]**	**Percent[3]**	**Percent Fixed[4]**
The total land use acres do not equal the total acres operated reported in the previous section on Acreage	71,957	41.6	89.2
Please enter a valid email address	53,667	31.0	61.9
The total acres operated by county exceeds the total acres operated for this operation	11,349	6.6	88.9
Please enter a response (no value/zero entered for number of men and women involved in decisions)	7,316	4.2	76.4
No value was entered for the total acres operated. Please enter your total acres operated or click Next if you do not operate any acres.	6,896	4.0	74.3
Please report number of acres harvested in 2017 (hay section)	6,595	3.8	83.7
Please report number of acres harvested in 2017 (field crops section)	5,830	3.4	83.2
Please report number of acres harvested in 2017 (vegetables section)	4,775	2.8	79.9
Your total acres operated cannot be less than 0	439	0.3	100.0
Please enter a 4-digit year prior to 2017 (year person began operating this/any operation-two questions)	340	0.2	89.1

4.4 Item Nonresponse

The item nonresponse for 26 variables was compared between web and mail and then web and CATI. Tables 12 and 13 show the item nonresponse rates for some of the variables. All variables for the web versus mail were statistically significant meaning the proportions are not equal. The web item nonresponse rate was between 0.5 to 11.5 percent. The mail item nonresponse rate was between 0.7 to 44.4 percent. Twenty-five of the twenty-six questions analyzed had lower item nonresponse for web compared to mail. However, the difference between the web and mail rates was less than four percent for nine of the ten demographic questions analyzed and greater than twenty percent for eleven of the fifteen screening questions analyzed. One of the reasons for the larger differences in the item nonresponse rates between the mail and web screening questions as noted in Figures 2 and 3 is that the commodity screening questions (same for the livestock) are all on one screen

[2] Total number of times warning messages triggered

[3] Percent out of all warning messages (respondent counts once per section). This will not add up to 100 percent since one of the critical error messages in the web form was not included in the table.

[4] Percent of the respondents for that warning message that fixed their response

for the web form as compared to the mail form where they are at the top of each section. Therefore, mail respondents who do not have a specific commodity tend to leave the entire section, including the screening question, blank.

Twenty-two questions were statistically significant for the web versus CATI comparison. The CATI item nonresponse was between 0.0 and 6.1 percent. Only nine of the questions had lower item nonresponse for web compared to CATI. However, the difference between the web and CATI was less than two percent for twenty-three of the questions. The plausible reason the item nonresponse rates are close between web and CATI is that an interviewer asks the questions and the CATI instrument has interactive features built in similar to the web such as skip patterns and warning screens.

Table 12: Item Nonresponse Rates: Web and Mail

Question Topic	Web Percent	Mail Percent	Web-Mail	Chi-Square P-value
Other livestock screening question	2.1	44.4	-42.3	<.0001
Aquaculture screening question	2.5	37.0	-34.5	<.0001
Woodland crops screening question	1.6	31.9	-30.3	<.0001
Access to the internet screening question	2.7	31.5	-28.8	<.0001
Year started operating any farm	5.6	16.2	-10.6	<.0001
Year started operating this operation	3.5	7.1	-3.7	<.0001
Household size	11.5	8.5	3.0	<.0001
Principal Occupation	1.6	3.7	-2.1	<.0001
Age	0.8	0.9	-0.1	<.0001

Table 13: Item Nonresponse Rates: Web and CATI

Question Topic	Web Percent	CATI Percent	Web-CATI	Chi-Square P-Value
Household size	11.5	4.5	7.1	<.0001
Age	0.8	2.2	-1.4	<.0001
Access to the internet screening question	2.7	4.0	-1.3	<.0001
Equine screening question	2.9	1.6	1.3	<.0001
Race	1.8	2.8	-1.1	<.0001
Aquaculture screening question	2.5	1.5	0.9	<.0001
Year started operating any farm	5.6	4.9	0.7	<.0001
Sex	0.5	0.0	0.5	<.0001
Year started operating this operation	3.5	3.8	-0.4	0.01

4.5 Modeling COA Response Mode

The characteristics related to web respondents were chosen for statistical comparisons and the models. Eleven variables in the COA were selected that were expected to show significant differences between web and mail. All differences between variables were statistically significant, meaning the proportions are likely not equal. A higher percentage of web respondents vs. mail respondents reported age less than 68; access to internet service; work other than farm or ranching; not retired; working one or more days off farm; "yes" to more than two crop/livestock screening questions; had higher burden (entered data for five or more sections); and had 3 or more household members. Web respondents do not

seem overburdened by the web form since they complete more sections, list more commodities, and have more household members (having to enter data for more persons in the household). Table 14 shows the rates for some of the characteristic variables.

Table 14: Characteristics of Web Responders versus Mail Responders

Question Topic	Category Description	Web Percent	Mail Percent	Web-Mail	Chi-Square P-value
Access to internet on operation or operator's residence	Yes	85.6	66.2	19.4	<.0001
	No	14.4	33.8	-19.4	
Age	less than 68	83.7	67.5	16.2	<.0001
	greater or equal to 68	16.3	32.5	-16.2	
Days worked off farm	None	29.8	44.5	-14.8	<.0001
	One or more days	70.2	55.5	14.8	
Household size	1 to 2	58.9	72.2	-13.3	<.0001
	greater or equal to 3	41.1	27.8	13.3	
Principal occupation	Farm or ranch work	36.2	51.4	-15.2	<.0001
	Work other than farm or ranching	63.8	48.6	15.2	
Retired	Retired	6.6	12.5	-5.9	<.0001
	Not Retired	93.4	87.5	5.9	
Burden measure	0 through 4 sections	47.9	63.9	-15.9	<.0001
	5 or more sections	52.1	36.1	15.9	
Crop and livestock commodities	0 through 2	71.9	80.9	-8.9	<.0001
	3 or more	28.1	19.1	8.9	

The bootstrap forest was used to further identify characteristics related to the mode. The response variable for the first model was web versus mail. The second model was web versus all modes (mail, CATI, CAPI). The models started with 32 predictor variables and ended up with 17 predictor variables. Both final models had an ROC value as shown in Figures 5 and 6 of 0.72 for the training and validation sets. Though higher values are preferred, the models were deemed as useful in this research.

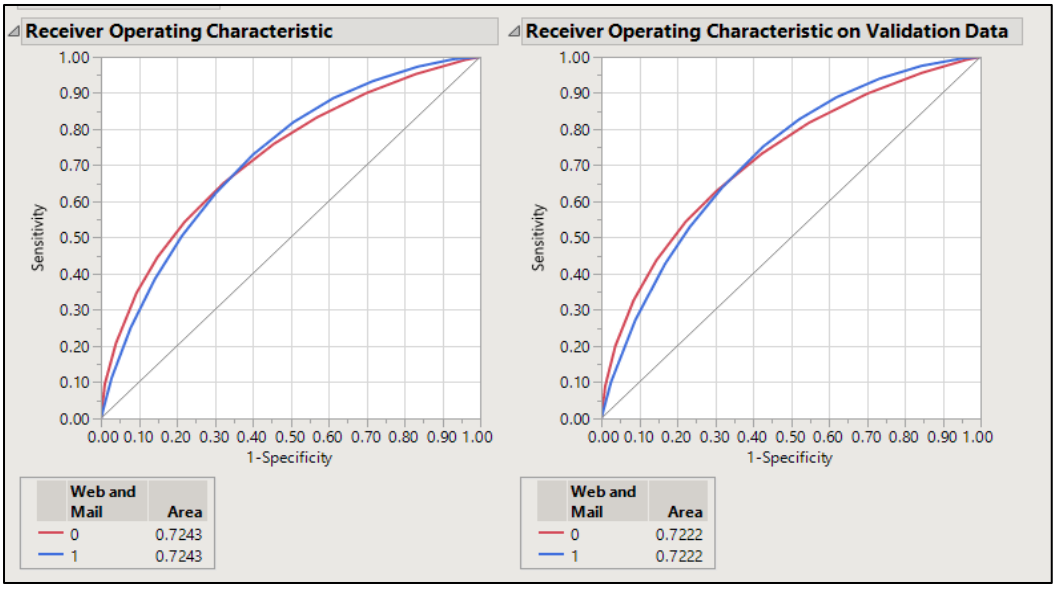

Figure 5: Web versus Mail ROC Training and Validation Final Model

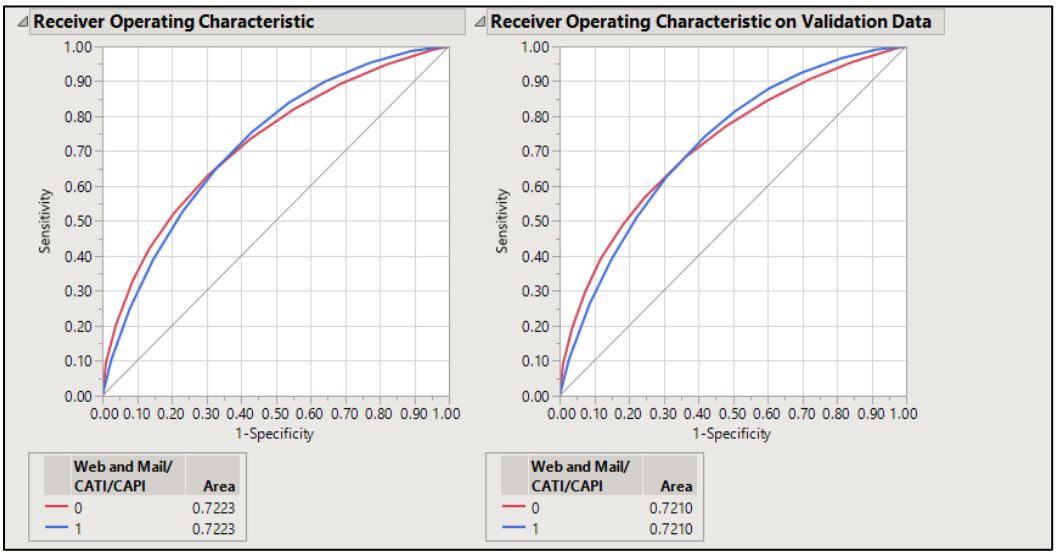

Figure 6: Web versus Mai/CATI/CAPI l ROC Training and Validation Final Model

The top contributors were similar between both models. This is expected since the number of CATI and CAPI cases is small (2 percent of all cases). The top five contributor variables for the web versus mail model were age, access to the internet, days worked off farm, household size, and retired as shown in Figure 7. The top five contributors for the web versus mail/CATI/CAPI model were age, access to the internet, principal occupation, days worked off farm, and household size as shown in Figure 8.

Term	Number of Splits	G^2	Portion
Age	3419	14232.9806	0.2661
Access to the Internet	1028	6455.96539	0.1207
Days Worked Off Farm	945	4815.40645	0.0900
Household Size	1921	3789.78761	0.0709
Retired	390	3289.25094	0.0615
Principal Occupation	957	3262.63191	0.0610
Farm Type	2073	2934.62438	0.0549
Internet Type: DSL	1045	2363.31391	0.0442
Internet Type: Mobile	1492	2335.82693	0.0437
Total Acres	1416	2164.2391	0.0405
Number of Completed Surveys Five Years	1614	1561.63798	0.0292
Race	48	1550.56582	0.0290
Region	2749	1450.86912	0.0271
Hybrid of Farm Type and NAICS Codes	1679	972.939136	0.0182
Sex	545	909.61458	0.0170
Number of Crop and Livestock Commodities	1718	791.675629	0.0148
Ethnicity	58	596.012113	0.0111

Figure 7: Web versus Mail Column Contributions

Column Contributions

Term	Number of Splits	G^2	Portion
Age	2537	12877.8392	0.2436
Access to the Internet	638	6701.7932	0.1268
Principal Occupation	600	5351.71229	0.1012
Days Worked Off Farm	708	4515.46904	0.0854
Household Size	1258	3284.89467	0.0621
Farm Type	1415	2796.15051	0.0529
Retired	329	2751.58191	0.0520
Internet Type: DSL	746	2692.62207	0.0509
Total Acres	921	2302.26303	0.0435
Internet Type: Mobile	1042	2119.38662	0.0401
Total Surveys Four Years	1287	1770.87672	0.0335
Region	1918	1287.7174	0.0244
Race	40	1280.00334	0.0242
Hybrid of Farm Type and NAICS Codes	1232	912.246328	0.0173
Burden Measure	1294	879.331689	0.0166
Sex	356	788.629899	0.0149
Ethnicity	32	554.54809	0.0105

Figure 8: Web versus Mail/CATI/CAPI Column Contributions

Decision trees were run for both final models to gain a better understanding of what is occurring within the model and the splits. Figures 9 and 10 show examples of the decision tree splits for both models. The models exhibit initial splits at age less than 68 and age greater than or equal to 68, and internet access yes and no.

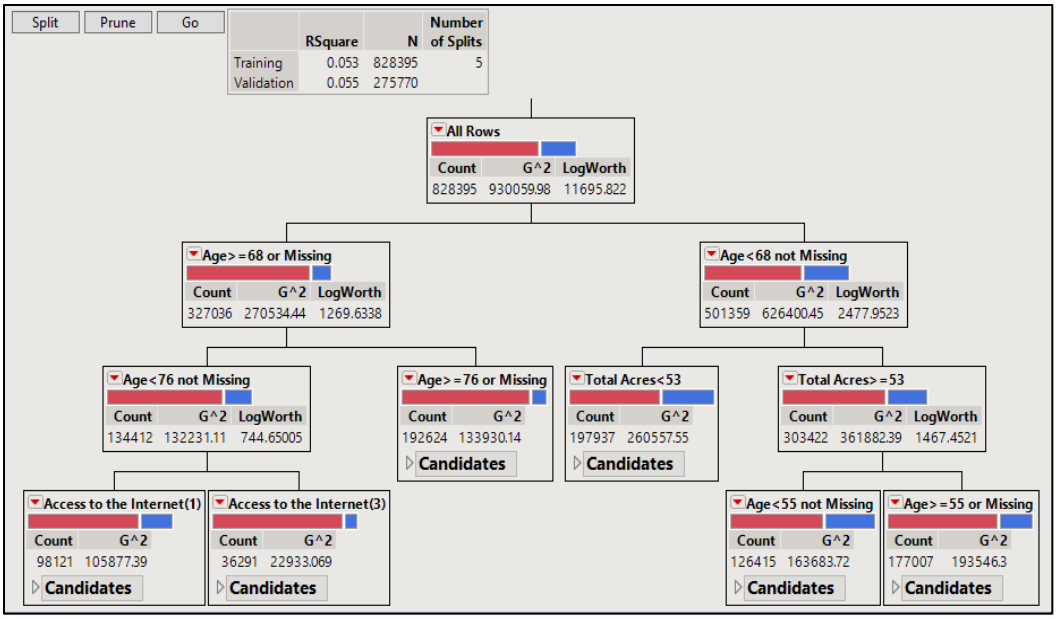

Figure 9: Web versus Mail Decision Tree

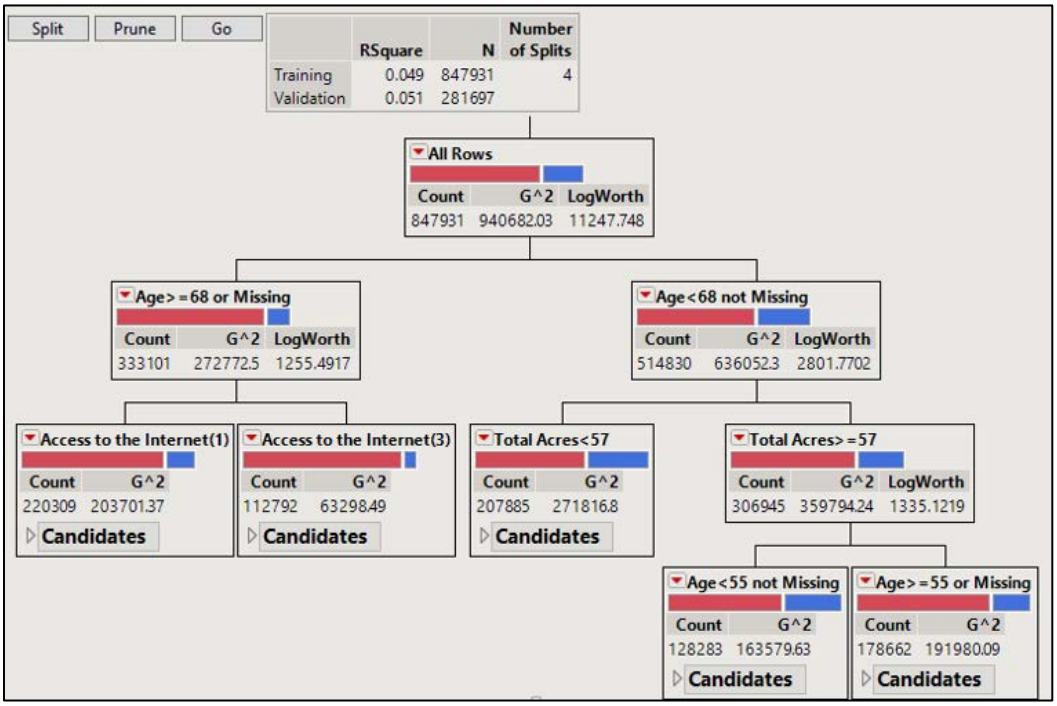

Figure 10: Web versus Mail/CATI/CAPI Decision Tree

Respondent level logistic regression models of web versus mail and web versus mail/CATI/CAPI were created using the highest contributors in the bootstrap forest models. Additional informative metrics were produced from the logistic regression models. The predictor variables chosen were the highest contributors to the bootstrap forest model. The covariates included age, household size, days worked off farm, retired/not retired, internet access, total acres, and having a cattle farm. For both models, the odds of a respondent who has internet access completing the form by web is 2 times more likely than the odds of a respondent completing the form by one of the other modes. Also, for both

models, the odds of a respondent who is less than 68 completing the form by web is 1.6 times more likely to occur than the odds of a respondent completing the form by one of the other modes.

5. Conclusions

Paradata plays a valuable role in helping survey organizations better understand their data collection processes to make informed decisions. There are several benefits to examining both device type and navigational paradata. For instance, paradata provides insight for a user-focused design, and can focus efforts on ensuring the web design is optimized for the most common devices and browsers. Additionally, questions where respondents often change answers could be investigated through cognitive testing and/or usability testing for future improvement. Also, additional review may be needed for the sections where the majority of respondents exited the web form early. There are challenges to working with paradata as discussed in the next section but the benefits far exceed them.

An advantage of respondents using the web form is the interactive features such as the warning messages and help buttons. Paradata showed that warning messages in the 2017 COA were effective in helping to minimize nonresponse and improve data quality. Approximately 46 percent of respondents triggered a warning message. Almost all warnings prompted majorities of respondents to correct their data.

The item nonresponse analysis also showed that web data collection exhibits lower item nonresponse compared to mail for the demographic and screening questions analyzed. However, the number of respondents that completed the form by each mode and their type of commodities (e.g., vegetables, field crops) must be considered when determining the impact that item nonresponse has on data quality.

Given analyses above and the high cooperation rates for both surveys analyzed, models to describe web respondents might be used to target those who normally wouldn't complete by web but are most similar to those who do. Some of the characteristics that differ between web and the other modes were identified by the bootstrap forest and logistic regression models. These included age, internet access, number of days worked off the farm, whether the operator was retired, number of household members, and the principal occupation of the operator. Additional studies could be done to better understand respondent preferences, as well as factors and obstacles to web completion. Also, making accessibility to the web instrument as easy as possible, improving communication on how to complete the Census online, and promoting web completion through effective marketing methods are also critical.

In conclusion, examining web paradata and the data quality across modes is important to evaluate data quality in the current and future surveys. The benefits are extensive including lower costs, improved data quality, and lower respondent burden. In the final section some challenges to working with the paradata are discussed.

6. Challenges

While paradata provided useful insights into our web data collection, there were several technical and logistical challenges to extracting and analyzing it. Some challenges included identifying the device and browser used, location of the break off, determining the question

where the respondent clicked the help button, and identifying the question where the respondent changed their answer. The paradata files contained a field called user agent string as shown in Figure 11. A user agent string identifies the browser, device, and operating system used by the respondent to access the web form. The difficulties were in identifying and extracting the information for analysis purposes. One of the websites used to parse out the information was *WhatIsMyBrowser.com*. Another challenge was determining the question where the respondent broke off or exited the web form early. One method is to pull the last question answered by using the time stamp. Another method which was used in this analysis is to pull the furthest question in the form completed since the respondent can go back and answer questions in previous sections. Another example of a challenge was extracting the location where the respondent clicked the help button. To evaluate the question where the respondent selected the help button, it was assumed that they clicked the button before they filled in their response to a question. A final example of a challenge was determining if the respondent changed an answer. There is a variable on the files that increments when the respondent changes their answer. However, a respondent can back space and re-enter the same answer. This will increment the variable on the file even though it is not a true change of answer. Therefore, the response to each question was used to identify if they changed their answer. These were just a few of the challenges working with the complex paradata files, but overcoming them allowed us to gain valuable insights.

User agent String (device=tablet, browser=internet explorer):
Mozilla/5.0 (Windows NT 6.3; WOW64; Trident/7.0; Touch; .NET4.0E; .NET4.0C; .NET CLR 3.5.30729; .NET CLR 2.0.50727; .NET CLR 3.0.30729; Tablet PC 2.0; F9J; CMNTDFJS; InfoPath.3; rv:11.0) like Gecko

Figure 11: Example of a User Agent String

References

Callegaro, M. (2013). Paradata in web surveys. In Kreuter, F. (Ed.), Improving surveys with paradata: Analytic uses of process information (pp. 261–280). Hoboken, NJ: Wiley.

COA (2017) Retrieved from https://www.nass.usda.gov/AgCensus/FAQ/2017/index.php.

Couper, M. P. (2000). Usability evaluation of computer-assisted survey instruments. Social Science Computer Review, 18, 384–396.

Crop/Stocks (2018) Retrieved from https://www.nass.usda.gov/Surveys/Guide_to_NASS_Surveys/Crops_Stocks/index.php.

Heerwegh, D. (2003). Explaining response latencies and changing answers using client-side paradata from a web survey. Social Science Computer Review, 21, 360–373.

Kaczmirek, L. (2008). Human-survey interaction. Usability and nonresponse in online surveys. Mannheim, DE: University of Mannheim.

Whatismybrowser.com (2010). Retrieved from https://developers.whatismybrowser.com.