Using Paradata to Inform Collection Instruments

Aneesah Williams, Chrishelle Lawrence, Andre Williams
U.S. Census Bureau, Governments Division, 4600 Silver Hill Road, Washington, DC 20233

Abstract
Paradata are data about the process by which survey data are collected. This paper explores the ways in which Governments (GOVS) Division can use paradata to improve our questionnaires. It looks at paradata from both web-collected surveys and paper survey forms. The 2011 Government Units Survey was collected using the Centurion web instrument, which enables the collection of paradata as respondents are completing their questionnaires online. Paradata such as keystrokes, time stamps of movement, and navigation patterns throughout the questionnaire were captured and analyzed. Some of the things analyzed include the frequency of response changes, time spent on each question, and survey break-off points. Additionally, paper responses to the 2011 Annual Survey of Public Pensions were examined to determine if there were questions that would benefit from further definitions or explanations, questions that were particularly troublesome for respondents, or break-off points on the questionnaire. Respondent call records and emails were also used to identify these problem questions.

Keywords: paradata, forms design

1. Introduction

Paradata are data about the survey collection process. This can include the time spent answering questions, the number of changes made to a response, the presence of error triggers, and countless other pieces of information. This paper explores the ways in which Governments Division (GOVS) of the U.S. Census Bureau can use paradata to improve survey questionnaires. In GOVS, paradata are gathered from web collection instruments—which capture responses, time stamps of movement, and navigation patterns—and paper questionnaires—which capture erasures, notes from the respondent, and break-off points. These paradata will be used to study how collection instruments can be informed and improved.

In this paper, paradata from the web-collection instrument for the 2011 Government Units Survey (GUS) and paper forms for the 2011 Annual Survey of Public Pensions (ASPP) were used. Paradata from these two surveys were analyzed to identify potential problem areas on the survey instruments. The background section of this paper will give a description of the surveys, basic terminology to be used throughout, and some of the challenges encountered during the analysis of the paradata. The remaining sections of the paper will discuss the details of these analyses and the results of each.

2. Background

2.1 Basic Survey Terminology

In this paper, we use the terminology described in chapter 1 of Kreuter (2013). Paradata are data about the process of generating the final product. They capture information about the data collection process. Paradata can be used to improve the quality of survey data by informing researchers of errors that occurred during survey production.

Additionally, paradata are increasingly being used to monitor the collection process and guide data collection procedures in a process known as adaptive design. This process includes the display and summarization of paradata in efforts to make adjustments to the data collection in near-real time. Some examples include survey mode switching, interviewer effort, and timing changes to non-response follow-up.

2.2 Governments Division Surveys

The Census of Governments is conducted every five years and has three components: Organization, Finance, and Employment. The Organization component uses the Government Units Survey (GUS) to identify the scope and nature of the nation's state and local governments, to determine the accuracy of the contact information, and to classify local government organizations, powers, and activities. Data collected include the number and function of the state and local governments throughout the U.S. These data help explain how each state is organized into different types of local governments (counties, cities, townships, and special districts) and how governments' responsibilities and authority vary from state to state, or within a state. The 2011 GUS was collected using both paper forms and the Centurion web instrument, a software that allows for the collection of paradata as respondents are completing their questionnaires. Forty-two percent of the GUS data were collected via Centurion, which provides time stamps for each selection a respondent makes throughout the survey questionnaire. The information received includes, but is not limited to, the unit identifier, the type of action taken (i.e., login, field change, error trigger, hyperlink selection, logout, etc.), the time each action is taken, the screen (or survey section) on which the action was taken, the web address if a link was selected, each field (question) name, and the value input in each field (question).

The 2011 Annual Survey of Public Pensions provides revenues, expenditures, financial assets and membership information for the defined benefit public pension systems for state- and locally-administered defined benefit systems. About 20 percent of the survey data were collected using the paper forms. These forms were examined for erasures, notes from the respondent, markings, and other forms of paradata.

2.3 Challenges

One of the more common challenges in working with paradata is the sheer amount of information that is collected. The Web-based paradata received from Centurion was a massive .xml file that would take hours to simply read in, sort, or even subset.

Another unexpected challenge was dealing with paradata that did not make sense. Time stamps seemed out of order or unreasonable. There were several respondents that appeared to complete the survey without ever logging in. Other respondents appeared to have submitted their survey before answering any questions, according to the time stamps. The latter, however, may have had a rational explanation—like the system time on the respondents computer being reset. Nevertheless, these IDs were excluded from several analyses due to the unknown nature of these anomalies.

We examined the paper forms looking for notes from respondents, miscellaneous marks made on the form, erasures, and other changes. However, we discovered that many of the marks made on the forms were not made by respondents, but by the analysts during data correction. Detecting which marks were from respondents and which were made by analysts took quite a bit of effort, as handwriting comparisons were made within and across forms. At the time the survey was being processed, some analysts erroneously moved responses from one question to another. Questionnaires with this type of marking were excluded from some of the analyses.

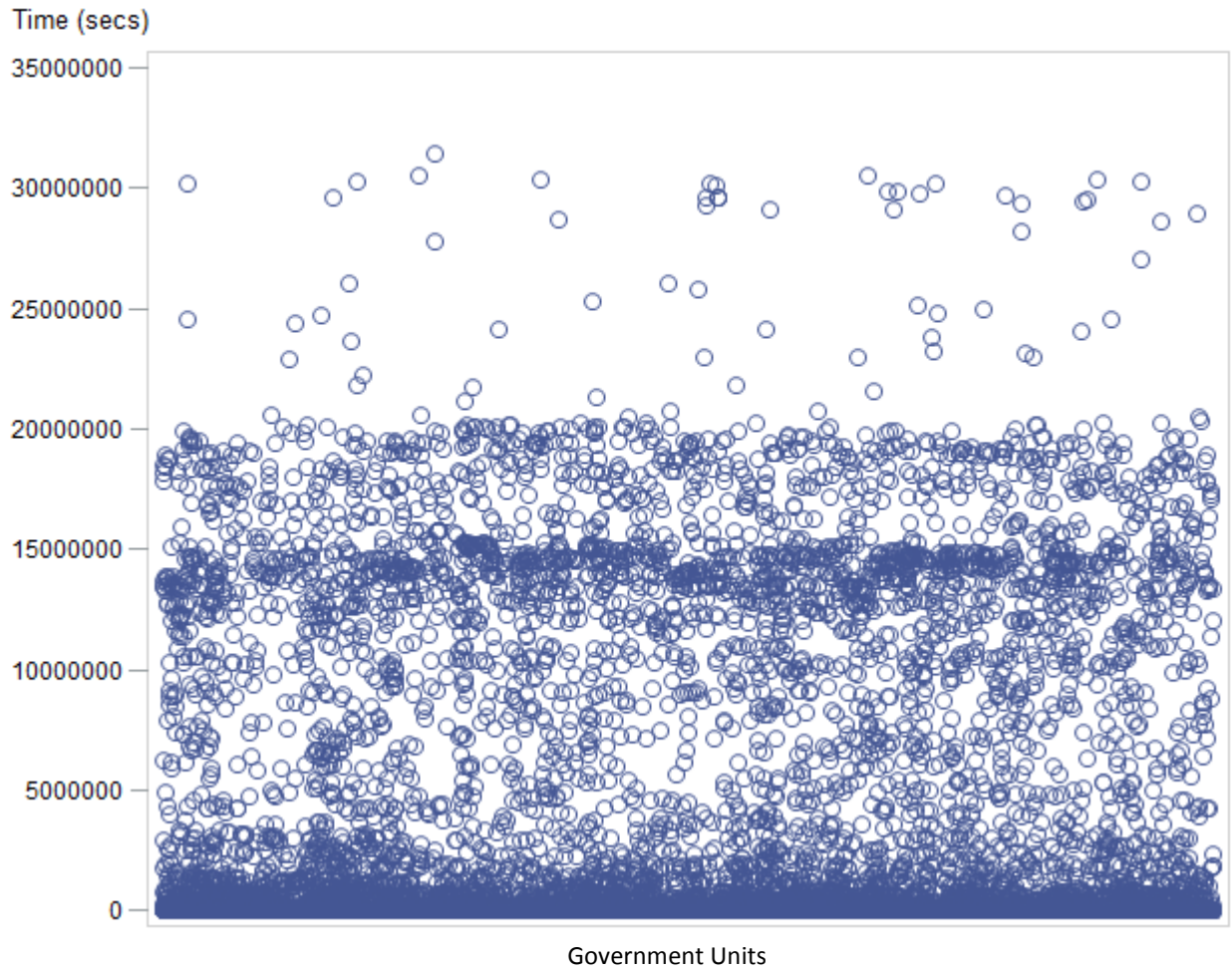
3. Web-based Paradata (GUS)

The Centurion paradata file was provided in an xml format. For each respondent, the events that took place (i.e., answering a question, opening a dashboard, logging in/out, etc.) during the completion of the survey questionnaire and their corresponding time stamps were captured. Additionally, information about the computing environment, the screen resolution, and the internet browser were captured.

3.1 Survey Completion Times

Overall, respondents took an average of three weeks to complete the questionnaire. Completion time was measured from the initial login to the questionnaire submission time (or to the last event for that respondent in the case where they did not submit). Figure 1 shows a scatter plot of completion times for the 25,071 Centurion respondent IDs with the exception of a few outliers. One respondent unit had an erroneous time stamp for its last event, which appeared to have occurred on December 31, 1999. Being entirely impossible, this unit was removed from the analysis of survey completion times. There were five outliers, IDs with a completion time that exceeded 75 weeks, which were removed from the analysis as well. Additionally, according to the paradata, one unit's only action was to finalize the survey without ever having logged in, thus was also excluded.

The average time to complete the survey, with above-mentioned units removed, was over two and a half weeks. The shortest completion time for a questionnaire was 8 seconds and the longest time was 52 weeks. To take into account the fact that respondents did not always complete the survey in one sitting, completion times were also calculated by cumulating the time spent on the survey only between logins and exits. In this respect, the average time to complete the survey fell to two weeks. Median completion times were 30 minutes overall and 26 minutes cumulative.



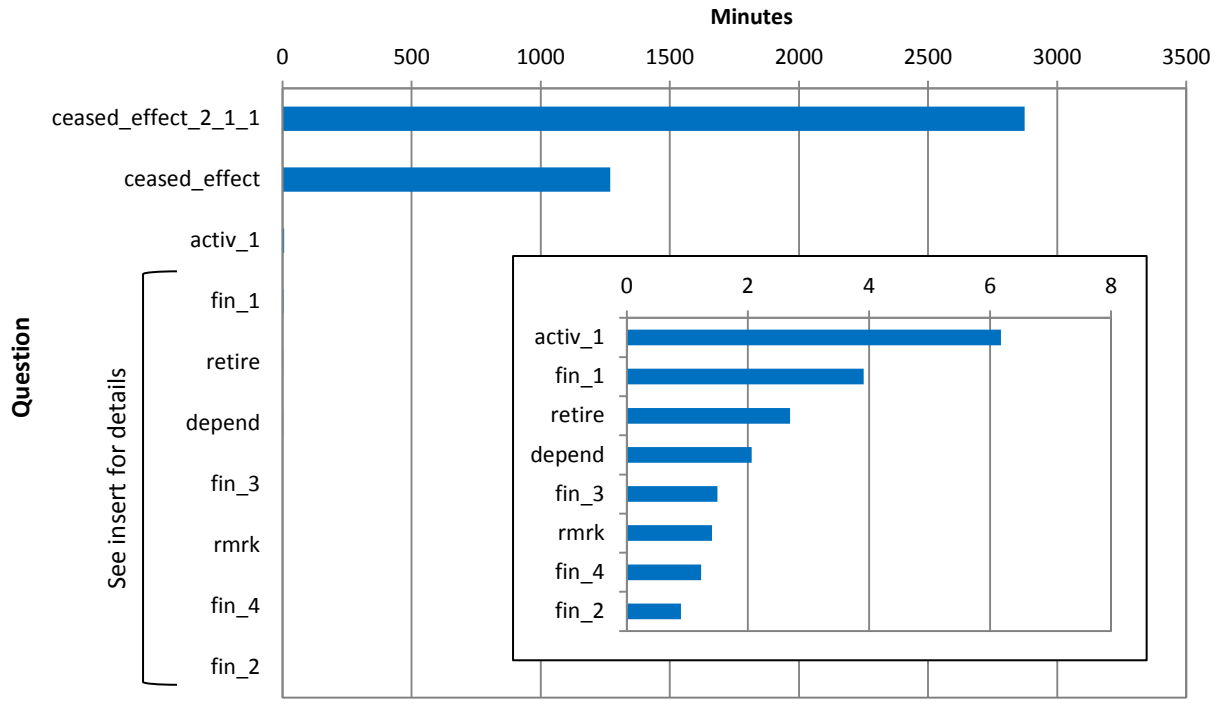
Source: U.S. Census Bureau, 2011 Government Units Survey Paradata File

Figure 1. Scatter Plot of Survey Completion Times

3.2 Question Completion Times

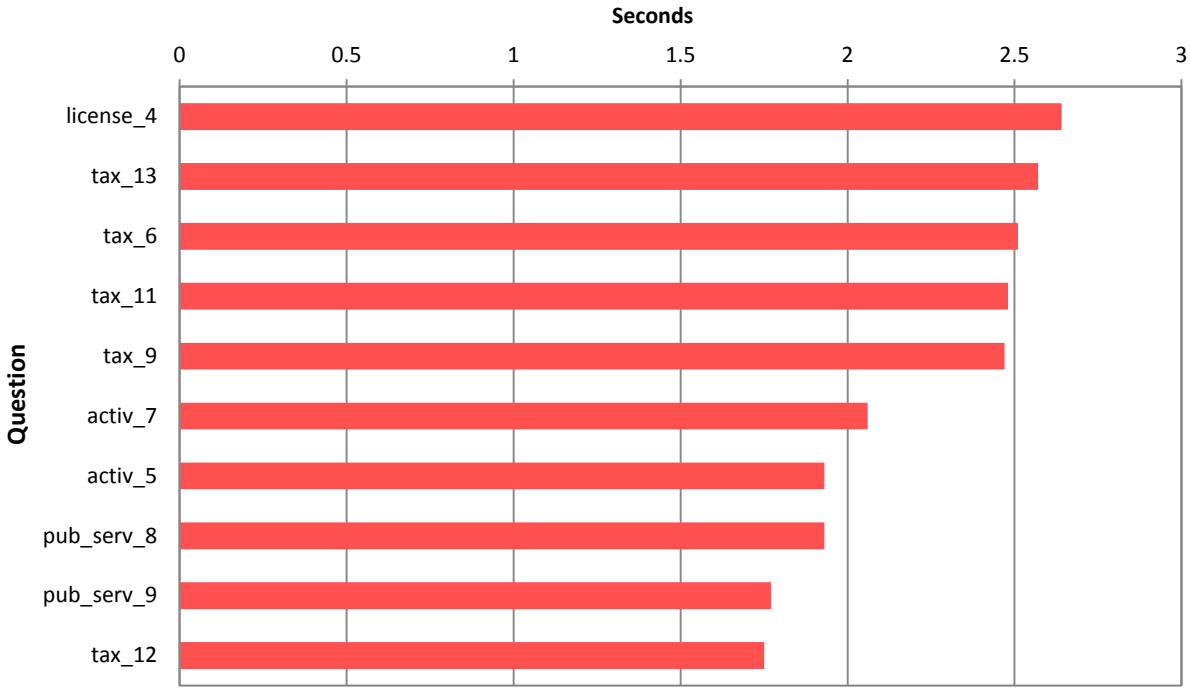
The time to complete each of the questions was also analyzed by measuring the event-to-event difference in time stamps. Overall, the average time to complete a question was 13 minutes. The median time was 3 seconds. Figure 2 shows the 10 longest average completion times for the questions on the survey and Figure 3 shows the 10 shortest average completion times. See Table 1 for a description of select survey questions, mentioned throughout.

The question with the longest average completion time was *CEASED_EFFECT_2_1_1*, indicating the effective month that the government ceased to exist, with over two days. *CEASED_EFFECT* took the next longest average time at 21 hours. Considering these were the first questions on the GUS survey, respondents that said a government had ceased to exist may have needed to research what the disincorporation date was. Also notable is that all four questions in the Finance section were among those with the longest 10 average completion times. These were the only write-in questions on the survey, asking for revenue, expenses, payroll, and debt of the government. *TAX_12* had the fastest average completion time of 1.7 seconds.



Source: U.S. Census Bureau, 2011 Government Units Survey Paradata File

Figure 2. Average Time Spent by Question (in minutes): The Longest 10 Times



Source: U.S. Census Bureau, 2011 Government Units Survey Paradata File

Figure 3. Average Time Spent by Question (in seconds): The Shortest 10 Times

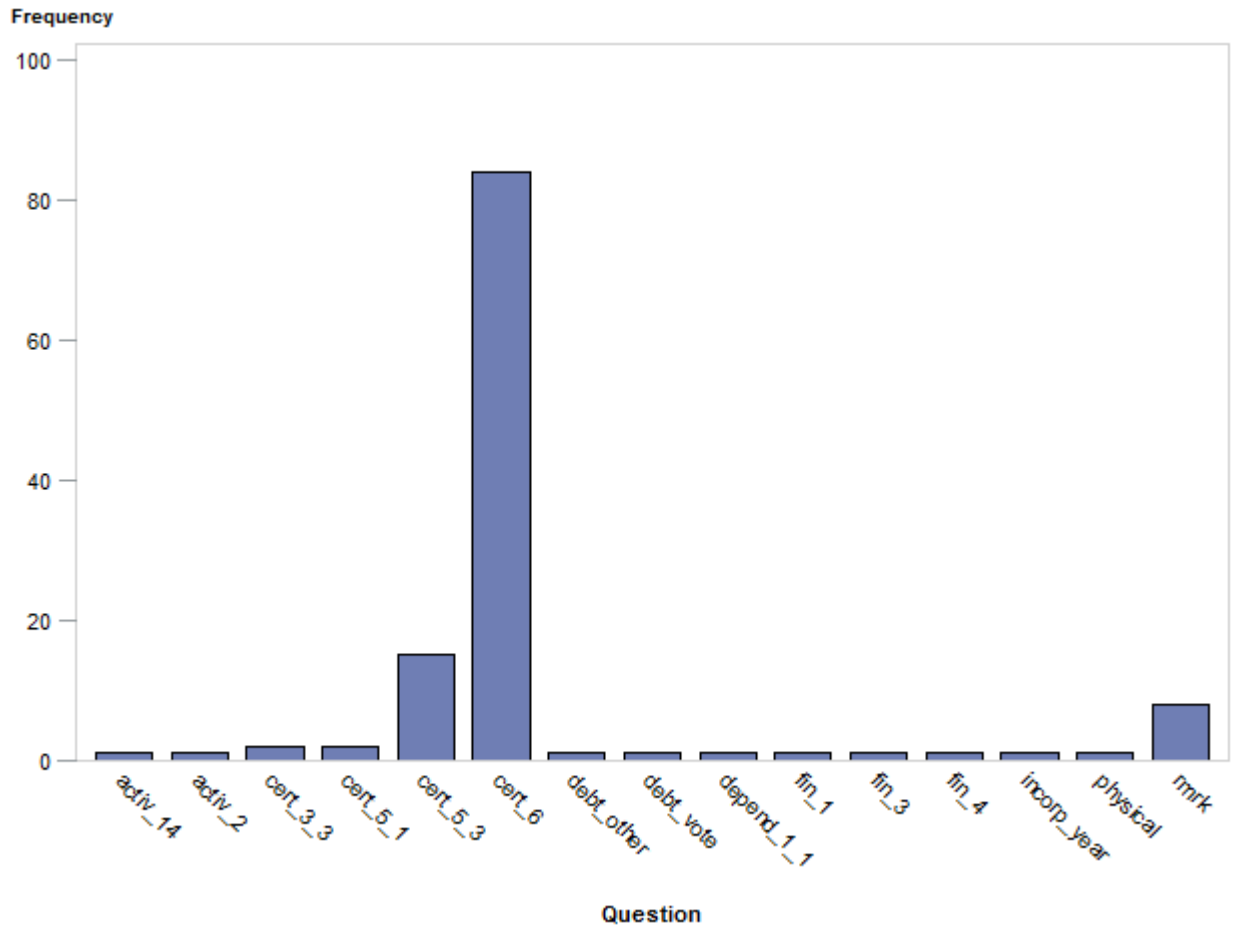
Table 1. Select Question Descriptions

Name	*Question*
ACTIV_1	Does your government operate a liquor store?
ACTIV_2	Is your government responsible for highways, streets, roads, alleys, bridges, tunnels, ferry boats, or related structures?
ACTIVE_5	Does your government operate an airport?
ACTIVE_7	Does your government operate a hospital?
ACTIV_14	Does your government keep separate records for the sewer system and the water supply?
CEASED_EFFECT	Was your government in existence on October 11, 2011?
CEASED_EFFECT_2_1_1	What was the effective month that government ceased to exist.
CERT_3_3	What is the contact person's telephone number? (last 4 digits)
CERT_5_1	What is the contact person's fax number? (area code)
CERT_5_3	What is the contact person's fax number? (last 4 digits)
CERT_6	What is the contact person's email address?
DEBT_OTHER	Is your government authorized to issue any other debt not specified above?
DEBT_VOTE	Does your government require voter approval to issue certain types of debt?
DEPEND	Is your government a fiscally dependent unit on another government, unit, agency, or office?
DEPEND_1_1	Is your government a fiscally dependent unit on a county government, unit, agency, or office?
FIN_1	How much revenue did your government receive in the last completed fiscal year?
FIN_2	How much did your government expend in the last completed fiscal year?
FIN_3	What was you government's annual gross payroll (before deductions) in the last completed fiscal year?
FIN_4	How much outstanding debt did your government have at the end of the last completed fiscal year?
LICENSE_4	Does your government have the authority to impose motor vehicles license fees?
PUB_SERV_8	Does your government provide any of the following types of library services – Academic libraries?
PUB_SERV_9	Does your government provide any of the following types of library services – Law libraries?
RETIRE	Do employees of your government participate in any retirement or pension plans?
RMRK	Please use this space for any explanations that may be important to understanding any of your responses.
TAX_6	Does your government have the authority to levy insurance premium sales tax?
TAX_9	Does your government have the authority to levy tobacco products sales tax?
TAX_11	Does your government have the authority to levy corporation net income tax?
TAX_12	Does your government have the authority to levy death and gift taxes?
TAX_13	Does your government have the authority to levy documentary and stock transfer tax?

3.3 Break-offs

Survey questionnaires were considered break-offs if they were not submitted by the respondent. The paradata file was analyzed in order to determine the most common break-off point across all survey responses. The goal was to pinpoint any question(s) that caused respondents to abandon completing the survey. One hundred and three (103) of the 122 respondent surveys that were not submitted, had been completed up through a question on the contact information and remarks page. This includes questions (shown in Figure 4) that begin with "*CERT*" and "*RMRK*." The fact that these questions were at the end of the questionnaire may indicate that the respondent merely forgot to submit the information. Perhaps it was their intention to return to the survey at a later time to submit it.

There were, however, earlier questions in the survey that were break-off points as well. Although only one respondent each was responsible for the remaining break-off points, some of those questions were: *ACTIV_14*; *ACTIV_2*; *DEBT_OTHER; DEBT_VOTE; DEPEND_1_1; FIN_1; FIN_3;* and *FIN_4*. See Table 1 for a description of these and other select questions.

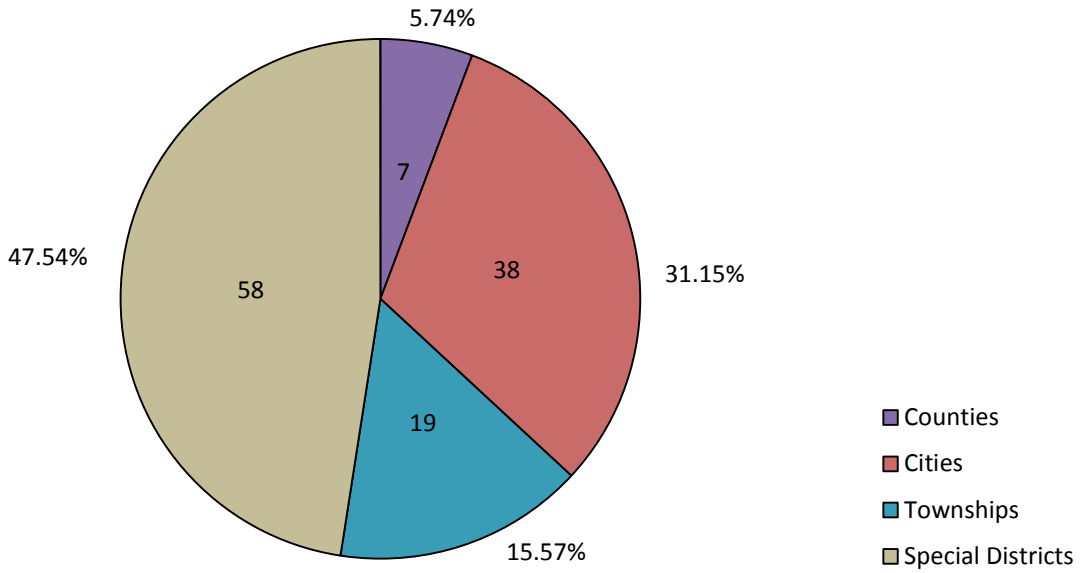


Source: U.S. Census Bureau, 2011 Government Units Survey Paradata File

Figure 4. Points of Survey Break-offs

When analyzing the break-off points by type of government, we found that Special District governments broke from the survey more than any other type of government (see Figure 5). Again, most of the break-off points were at the end of the survey where contact information was collected. However, special district units also broke-off in the section asking debt questions and even prior to that in the background

section. Questions in the background section asked whether the unit was dependent on another government entity and what year the government was incorporated. While true that there are more special districts than any other types of government, their governments are usually much less complex than cities and counties.

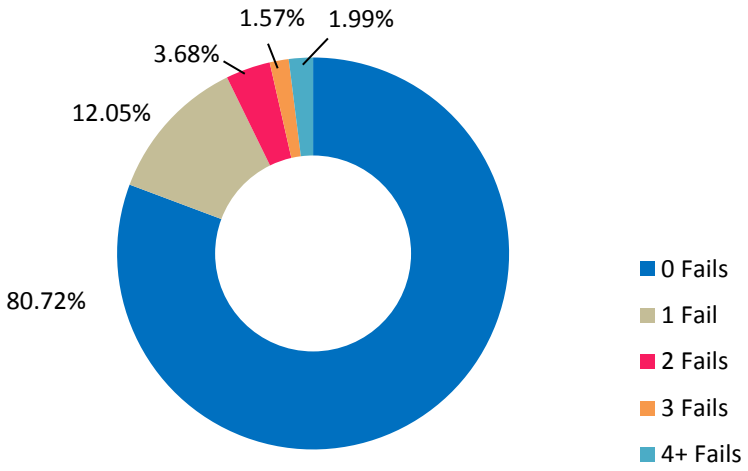


Source: U.S. Census Bureau, 2011 Government Units Survey Paradata File

Figure 5. Break-offs by Type of Government

3.4 Failed Logins

Respondents generally were able to log into the survey without issue, as shown in Figure 6. Nearly 81 percent of respondents never experienced a login failure. However, the greatest number of failed logins for a given respondent was 20. Over 3,000 respondents (12 percent) had one failed login during the course of their questionnaire completion.



Source: U.S. Census Bureau, 2011 Government Units Survey Paradata File

Figure 6. Number and Percentage of Failed Logins

3.5 Response Changes

We observed the number of times an answer was changed by a respondent. The average number of times an answer changed overall was less than one. The largest number of changes for one question was 119. That respondent changed the answer to *TAX_7* (Does your government have the authority to levy pari-mutuels sales tax?) by toggling back and forth between yes and no for 5 seconds straight. This was not the only instance, as this respondent also toggled between yes and no on questions *ACTIV_8* (Does your government own a gas utility?) and *ACTIV_9* (Does your government own an electric utility?) which were changed 75 and 49 times, respectively, over the course of 17 seconds.

4. Paper Paradata (ASPP)

Paper forms were examined to determine if there was any information that would provide insight on potential areas of improvement for data collection. The 2011 ASPP form collects the following information on defined benefit plans: Plan Information, Membership and Benefits, Receipts/Payments, and Holdings and Investments. Some of the things analyzed include written calculations on the form, explanations of answers, written refusals, erasures, crossed out data, and other markings. We examined forms for these types of paradata and attempted to quantify instances of their occurrences. Among the paper forms that were examined, there were no instances of a respondent breaking off from the questionnaire.

4.1 Types of Paradata

Of the 292 forms reviewed, 90 forms (30.8 percent) had paradata that fell into the following seven categories:

- *Gave percentage and amount*
  On some forms, dollar amounts were asked for and provided for three categories. In addition, percentages of the total of the three dollar amounts were written near the response box or in the margins. This happened twice for the benefits payments question (Z13, Z14, and Z15) and once for amount paid to beneficiaries receiving periodic benefit payments (Z08, Z09, and Z10) on three separate forms.

- *Inconsistent data*
  Most of the inconsistent data came from items Z08 (monies paid to members retired on account of age or service), Z10 (monies paid to members retired due to disability), and Z11 (monies paid to former or active members for withdrawals and one-time payments). Although the corresponding number of payees (items Z03-Z05) were given, these items were left blank. It is important to note that Z08, Z10, and Z11 were all removed from the form for survey years 2012 and beyond. New questions were added that asked about payments to beneficiaries.

  The remaining inconsistencies came from item V19 (employer normal cost as a percentage of covered payroll). Respondents have the option of providing the actual dollar amount for employer normal cost *or* the percentage of covered payroll. If the respondent answered both questions, it was considered inconsistent. All of the inconsistent data came from state units, as this question was only on the State administered survey form.
- *Markings*
  Common markings were arrows, brackets, underlines, and circles. The only expected marking would be an asterisk (*) to indicate that an item was estimated. Asterisks were written near eight items across four different forms.
- *No details provided*
  When a respondent was unable to break down large amounts into individual items, it was classified as ‘no details provided’. This occurred when a total amount was supplied by the respondent but detail amounts were not.
- *Remarks*
  Remarks were any information provided in the designated Remarks section of the form. Typically, respondents used this area to further explain an answer to a previous survey question.
- *Response change*
  A response change occurred when all or part of the answer was scratched out or White-out was used to cover an answer.
- *Wrote additional explanation*
  Additional explanations occurred when text was written somewhere on the form other than an answer box. Respondents occasionally wrote between a question and its answer box, in addition to writing their answer in the box.

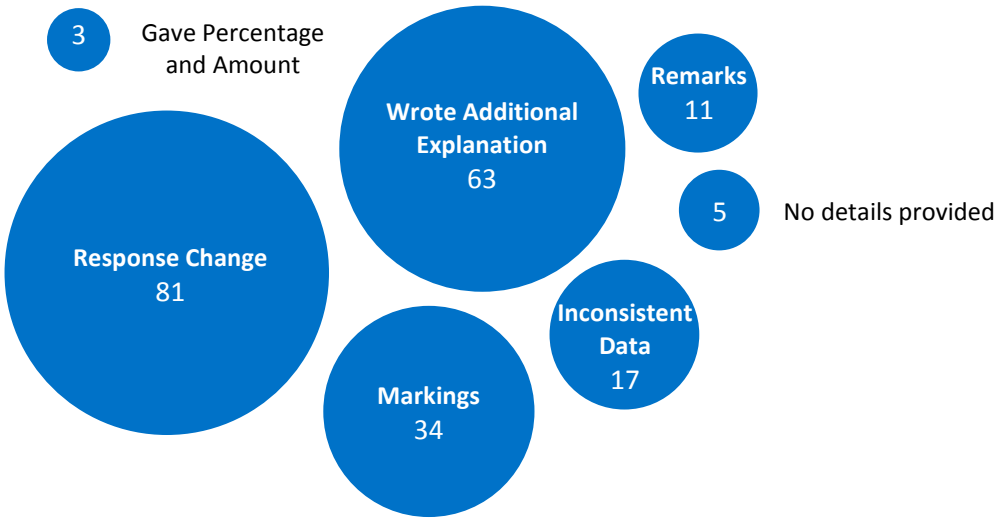
A list of select survey items from the 2011 ASPP, which are noted in the paradata analyses, is given in Table 2.

Table 2. Select Survey Item Descriptions

Item	*Description*
V19	Percentage of covered payroll estimate of employer normal cost
X01	Employee contributions
X04	Employer (government) contributions from parent local governments
X08*	Total earnings on investments (sum of Z98*, Z71, Z72, Z73)
X11	Total benefit payments (sum of Z13, Z14, Z15, Z16)
X21	Total cash and short-term investments (sum of Z88, Z87, Z68)
X30	Total federal government securities (sum of Z89, X33)
X33	Federal agency securities
X35	State and local government securities
X44	Total other securities (sum of Z84, X35, Z70, Z83)
Z02	Number of inactive members of retirement system
Z03	Number of payees: former active members of system, retired on account of age or service
Z04	Number of payees: former active members of system, retired on account of disability
Z05	Number of payees: survivors of deceased former active members
Z08	Amount paid during month: former active members of system, retired on account of age or service
Z09	Amount paid during month: former active members of system, retired on account of disability
Z10	Amount paid during month: survivors of deceased former active members
Z11	Amount paid during month: withdrawals and other one-time payments (other than loans) made to present or former members of system
Z13	Retirement benefit payments during fiscal year
Z14	Disability benefit payments during fiscal year
Z15	Survivor benefit payments during fiscal year
Z16	Other benefits
Z62	Federally-sponsored agencies' corporate bonds
Z63	Other corporate bonds
Z70	Foreign and international securities
Z71	Interest earnings on investments
Z72	Dividend earnings on investments
Z73	Other investment earnings
Z77	Total corporate bonds (sum of Z62, Z63)
Z78	Corporate stocks
Z81	Total cash and security holdings of public employee retirement system (sum of X21, X30, Z77, Z78, X42, X44, Z82)
Z83	Other securities
Z84	Investments held in trust by other agencies
Z87	Time or savings deposits
Z88	Cash on hand and demand deposits
Z89	Federal treasury securities
Z98*	Rentals from the state government

* Items found only on the F-12 State-administered ASPP forms.

Figure 7 shows the distribution of paradata found. The most popular category was “response change”, followed by “wrote additional explanation.”



Source: U.S. Census Bureau, 2011 Annual Survey of Public Pensions

Figure 7. Frequency of Types of Paradata

4.2 Data Value Changes

When looking at the largest source of paradata, “response change,” most of the changes were data value changes. Of the 90 forms reviewed, 36 forms (40 percent) had at least one data value change. On average, there were 2.4 changes per form, with the most being 11 changes on one form. When considering the type of value change (changed from zero, changed to zero, non-zero value-to-value change), most were unable to be determined (77.8 percent) because the responses were illegible having been removed with White-out or scratched out with ink. The next largest group was the non-zero value-to-value changes (10 percent). The survey item changed most frequently (7 times) was Z81, total cash and security holdings. On all but one of the forms with a Z81 change, at least one of the detail items summing to Z81 was changed as well.

Table 3 shows the average, minimum, and maximum difference of data values for all survey items where the old and new values were able to be determined. The largest differences occurred with X44 (total other securities). Its maximum difference was $-66,006,964, where the respondent initially did not sum the details to get the total. Respondents neglecting to sum detail survey items up to the X44 total survey item occurred on two other forms. The smallest data value change was -6 for Z05 (survivors of deceased former active members receiving periodic benefit payments).

Table 3. Data Value Changes by Item

Item	*Frequency*	*Difference in Data Values*		
		Average	*Minimum*	*Maximum*
X04	1	-14,314	-14,314	-14,314
X08	2	120,896	-66,938	308,729
X21	1	-50	-50	-50
X44	4	-19,218,614	9,119	-66,006,964
Z03	1	62	62	62
Z05	1	-6	-6	-6
Z08	1	105,639	105,639	105,639
Z10	1	54,527	54,527	54,527
Z14	1	1,386,611	1,386,611	1,386,611
Z63	1	1,000	1,000	1,000
Z73	1	66,938	66,938	66,938
Z81	2	-1,331,639	-9,179	-2,654,098
Z83	1	9,119	9,119	9,119
Z84	1	150,559	150,559	150,559

Source: U.S. Census Bureau, 2011 Annual Survey of Public Pensions

Table 4 shows the survey item by the number of forms where the survey item was changed. Thirty-seven out of the 50 item codes experienced a data value change on at least one form. It is important to note that the items changed most frequently (Z81, X21, X44) are totals.

Table 4. Frequency of Data Value Changes

Number of Forms	*Item(s)*
7	Z81
6	X21, X44
4	X04, X08, Z08, Z10, Z88
3	Z02, Z05, Z63, Z77, Z78
2	X01, X11, X33, Z03, Z62, Z70, Z83, Z89, Z96/Z91
1	X05, X30, Z01, Z04, Z06, Z09, Z13, Z14, Z16, Z68, Z71, Z72, Z73, Z84, Z87

Source: U.S. Census Bureau, 2011 Annual Survey of Public Pensions

5. Summary

There are many pieces of information that can be found in the paradata of survey questionnaires. So much can be learned from observing the respondent's movements through the survey instrument when viewing web-based paradata. Similarly, notes and markings on paper instruments can provide insight to a respondent's motivation and intent when answering survey questions. Analyzing the paper paradata gave us insight into the data review process issues.

What the above analyses showed most is that there are still short comings to paradata. Web-based paradata can have issues with their time stamps, missing observations, and other computer related irregularities. Paper paradata too have issues, mostly due to human error. Differentiating between respondent marks and analyst marks was a particularly challenging task in this research.

Acknowledgements

The authors would like to acknowledge the subject matter experts who assisted with this research. Rachelle Reeder and Ceci Villa Ross provided invaluable information about the Government Units Survey and Annual Survey of Public Pensions, as well their time. We would also like to thank Carma Hogue and Suzanne Dorinski for their assistance in retrieving, understanding and working with the Centurion paradata.

Reference

Kreuter, Frauke, ed. (2013). *Improving Surveys with Paradata: Analytic Uses of Process Information*. Hoboken, NJ: John Wiley & Sons, Inc.