Paper #AD08

A Picture is Worth a Thousand Data Points: Using SAS Arrays, LAG Functions and ODS to Create a Flexible Survey Interviewer Tracking Chart

Joe Murphy, Adam Safir, Susan Myers, and Christy Hottinger, RTI International[1], Research Triangle Park, NC

ABSTRACT

Advances in technology have made available a staggering amount of information on the survey data collection process for use in tracking field interviewer (FI) behavior. The data can be reviewed to detect inefficiencies or protocol violations that may reduce data quality. The challenge lies in how to harvest these rich data sources to produce valuable tools for those who monitor the data. This paper presents a SAS technique developed by RTI International to compile raw record-of-calls data from individual field interviewers and display them in an interactive and informative format for data quality reviewers and field supervisors. Record-of-calls data contain the date, time, case ID, interviewer ID, and disposition code for every call made to a household selected for a survey. Our process reads the raw data into SAS, organizes them into the necessary output format using arrays and the LAG function, and uses the SAS Output Delivery System (ODS) to generate a Microsoft Excel-based report. With the report, a reviewer is not only able to trace the pattern of calls made by an interviewer to a single household on a single day, but can also get an overall picture of the progress of the field activity and make comparisons across interviewers and regions. In this paper, we present our process and code for creating the data review tool, and discuss the value of its implementation. This paper is geared towards those who review data or design review tools and is intended for SAS users of all skill levels.

INTRODUCTION

Over the last several years, the amount of data available to survey researchers on the data collection process has grown vastly. With the adoption of Computer Assisted Interviewing (CAI) technology on most major field surveys, researchers can now compile and analyze "process data," or data about the interviewing process, that was not available in the past. For instance, one can analyze the work patterns of field interviewers including the times and days they make their calls, the length of their interviews (or non-interview calls), frequency of visits to particular households, and outcomes or dispositions of the individuals calls. While these data undoubtedly have enormous potential to inform field supervisors and reviewers of the data about potential sources of interviewer-related inefficiencies and potential protocol violations, researchers are just beginning to determine the best way to sort through the numerous fields of data to harvest actionable insights.

SAS software provides a means to transform raw data on the data collection process into a valuable tool for survey supervisors and researchers. This paper presents a process developed by RTI International to organize raw record-of-calls data into a summary chart that can be used to identify inefficient or potentially fraudulent behavior by field interviewers. The process utilizes SAS arrays, date functions, and the Output Delivery System (ODS) to produce a Microsoft Excel-based report depicting interviewer behavior from the individual call attempt level to the household and interviewer levels. The report also allows for comparison of behaviors across interviewers. By transforming the data into a more readable format, the tool greatly reduces the time needed by supervisors to review interviewers' work and increases their power to detect heretofore unseen patterns of behavior. This power allows the supervisors to act upon evidence for the purpose of increasing efficiency in the data collection process and overall survey data quality.

BACKGROUND

Survey process data, sometimes referred to as administrative data, or paradata, are a form of metadata which provide information on the data collection process itself. Process data are extensive and varied, although they can generally be classified into one of two types:

- macro process data, the more common of the two, which take the form of global process summaries, such as overall response or coverage rates; and

- micro process data, the less well-known type, which provide detailed information on each interview case, such as the number of attempts made before first contact, what language the interview was conducted in, or the tenure of the interviewer who completed the case.

[1] RTI International is a trade name of Research Triangle Institute

In this paper, we will use the generic term process data to refer exclusively to micro process data.

Although process data have been generated as a by-product of survey administration since the earliest data collection efforts were undertaken, recent advances in interviewing technologies have markedly increased the ability of data producers to efficiently store, manipulate, and utilize process data for a wide range of systems improvement. In particular, process data present opportunities for systems improvement in design, operations, and quality control in face-to-face, telephone, and web-surveys.

In the realm of web-survey design, client side process data have enabled researchers to observe how respondents proceed through web questionnaires by recording user actions paired with precision time-stamps. Researchers use these process data to study methodological issues such as satisficing, navigational error rates, attitude strength, and respondent burden (Heerwegh, 2002 & 2004). Additionally, Bosnjak and Tuten (2001) relied upon process data to construct a typology of response behavior in web surveys.

Operationally, process data have been used to optimize contact attempt protocols for telephone and face-to-face surveys (Groves, et al, 2003; Kalsbeek, et al, 1994; Kulka and Weeks, 1988; Purdon, 1999; Weeks, et al, 1987). Other operational uses of process data have centered on understanding how interviewers allocate their efforts and whether the corresponding level of effort is related to survey measures of interest (Lipps and Benson, 2005; Wang, et al, 2005).

Finally, process data represent an invaluable tool for examining the quality of data beyond that which is communicated through response rates and sampling error, making possible a more comprehensive and independent evaluation of non-sampling error in survey estimates of interest (Safir, et al, 2001). Using process data, data users can construct quality assessments that are conditional upon their own unique analysis constraints (Scheuren, 2001).

PROBLEM

Survey data are used extensively by private and federal entities to determine the need for research, policy, and funding. Inaccuracies in these data either from poor data quality due to inefficient data collection, interviewer protocol violations, or falsification can lead to erroneous and costly research conclusions as well as misguided policy. Although falsification in survey research is believed to be a rare event, every instance of fraud or protocol violation damages the integrity and validity of the data collected. Thus survey researchers must have a comprehensive verification system in place to minimize this damage.

It would be unrealistic to expect survey researchers to verify one-hundred percent of the data collected due to cost and time restraints on research projects. What is both realistic and needed is a timely detection of low quality work and suspicious or fraudulent behavior among data collectors. Timely detection allows for focused verification and remediation of any low quality or invalid data within the time and cost restraints of the project, thereby protecting the integrity of the research findings.

Data quality reviewers at RTI have asserted that there exists a meaningful association between difficult-to-complete interviews and falsified cases. Current data mining techniques used at RTI have supported this assertion. This suggests that the decision to falsify is made after the interviewer has experienced difficulty in contacting the subject. However, completing an expert review of all the ROC data collected within a survey period would be very challenging given the sheer volume of records available.

SAS SOLUTION

Using SAS, we were able to create a usable and informative interviewer tracking chart for data reviewers and field supervisors to more efficiently analyze survey process data. The resulting report enables the reviewer to quickly and easily track trends in the data at the individual interviewer, case, and project level; allowing for quick detection of a known or emerging pattern of suspicious work behavior that may need closer examination.

The process data used for this paper are record-of-calls data recorded by field interviewers (FI) on a hand-held electronic device, the iPaq. The iPaq is loaded with case information assigned to a specific FI. It has an identification number for the FI (FIID) as well as a list of household IDs (DUID) assigned to the FI.

A screening of the household is conducted first to see if anyone in the residence is selected for an interview. To monitor the progress of cases, FIs enter event codes each time they attempt to contact a dwelling unit. The iPaq logs the date and time each event is entered. The number of events recorded for each household may range from the minimum two, initialization and complete screening code, to forty or more. The cases with a high volume of events are typical when respondents are not home. The number of events per case is not sufficient in itself in

assessing whether it is acceptable. Cases are worked on a quarterly basis, so an FI has three months to complete a screening and possible subsequent interview.

Record-of-calls data are transmitted nightly to RTI International where they are stored in a MS SQL Server database. The volume is quite large and depending on the size of the study can total over one million records per year. The iPaq also has a note field available for the FI to document in more detail what is happening with a case. Due to the volume of data, our SAS dataset does not incorporate the comment, but it is available for analysts if needed.

Our process reads the raw iPaq data into SAS, manipulates the layout, exports using the ODS, and creates a Microsoft Excel chart for data reviewers. The data read into SAS follow the format shown below.

fiid	duid	date_time	code
999123	IL99999903	01JUL2004:11:35:00	1
999123	IL99999903	03JUL2004:12:21:00	1
999123	IL99999903	03JUL2004:12:50:00	1
999123	IL99999903	05JUL2004:13:53:00	1
999123	IL99999914	02JUL2004:12:42:00	1
999123	IL99999914	03JUL2004:12:29:00	1
999123	IL99999914	06JUL2004:13:56:00	4
999123	IL99999934	01JUL2004:12:53:00	4
999123	IL99999945	01JUL2004:12:35:00	1
999123	IL99999945	05JUL2004:14:03:00	1
999123	IL99999955	01JUL2004:12:43:00	2
999123	IL99999965	01JUL2004:12:36:00	1
999123	IL99999965	03JUL2004:12:55:00	1
999123	IL99999965	06JUL2004:14:28:00	1
999123	IL99999996	01JUL2004:12:16:00	1
999123	IL99999996	05JUL2004:14:30:00	1
999123	IL99998817	01JUL2004:12:31:00	1
999123	IL99998817	03JUL2004:11:08:00	3
999123	IL99998817	03JUL2004:11:08:00	3
999123	IL99998817	03JUL2004:22:08:00	4
999123	IL99998838	01JUL2004:12:24:00	1
999123	IL99998838	03JUL2004:11:15:00	1
999123	IL99998838	05JUL2004:14:30:00	1
999178	IL99998848	01JUL2004:12:30:00	2
999178	IL99998848	03JUL2004:11:13:00	1
999178	IL99998848	05JUL2004:14:34:00	1
999178	IL99998858	01JUL2004:13:15:00	1
999178	IL99998858	01JUL2004:22:25:00	2
999178	IL99998858	02JUL2004:14:01:00	4
999178	IL99998869	01JUL2004:13:27:00	1
999178	IL99998869	02JUL2004:12:58:00	1
999178	IL99998869	02JUL2004:12:58:00	3
999178	IL99998869	02JUL2004:15:51:00	4
999178	IL99998879	01JUL2004:13:07:00	4
999178	IL99998889	01JUL2004:13:47:00	1
999178	IL99998889	01JUL2004:13:47:00	3
999178	IL99998889	03JUL2004:11:17:00	3

Each record represents a call transaction. The first column (fiid) gives the ID number of the FI entering the record of the call, the second column (duid) gives the ID number of the dwelling unit (DU) to which the call was made. The first two characters of duid indicate the state (in this example, Illinois). The third column (date_time) gives the date and time of the call in the DATETIME20. format. The fourth column (code) gives the value of the disposition code of the call. In this example, the following disposition codes are used:

- 1: no one was home when the FI visited the household
- 2: some one was home, but no one was available for the interview
- 3: the householder refused to complete the interview
- 4: the interview was completed

Equipped with this file, a reviewer may be able to sort through the records to piece together the pattern of work of interviewers. For instance, a reviewer might note that FI 999123 visited several households on July 1 but only

completed 1 interview on that day. Further the reviewer might see that each of these calls was made before 4:00 pm, when the chance of finding someone at home is generally lower than after 4:00 pm. The reviewer could then go through the record-of-calls to find similar cases worked by that or other FIs to see if this potentially inefficient pattern of calling was repeated on other cases. This process would be very time consuming and is not very intuitive. This is especially true for very large data collection efforts. One effort conducted by RTI includes almost 200,000 households each year. The record-of-calls file for this study contains more than 1,000,000 records a year.

To provide a tool that cuts down on review time and reveals heretofore unseen patterns, we developed the annotated code presented here in several steps.

STEP 1 – READ IN THE RAW DATA

The first step reads in the raw ROC data and uses date functions to create variables for the date, year, quarter of the year, day of the week, and hour of the day for each call.

```
data roc;
      informat fiid $6. duid $10. date_time datetime20. code 2.;
      format date_time datetime20.;
      infile 'c:\temp\roc.txt';
      input fiid duid date_time code;
      date=datepart(date_time);
      year=year(date);
      qtr=qtr(date);
      day=weekday(date);
      hour=hour(date_time);
```

STEP 2 – CREATE AN AFTERNOON/EVENING INDICATOR AND LIMIT THE DATA TO A SINGLE STATE

Next, a variable is created to show whether the call occurred after 4:00pm and the data are limited to a single state (Illinois) and a single week (July 1, 2004 to July 7, 2004).

```
      if hour>=16 then after4=1;
            else after4=0;
      cutoff=input('07JUL2004',date9.);
      if qtr=3 and substr(duid,1,2)='IL' and date<=cutoff;
```

STEP 3 – SORT THE DATA

The data are sorted by interviewer ID, household ID, date, time, and disposition code.

```
proc sort data=roc;
      by fiid duid date_time code;
```

STEP 4 – CREATE ONE RECORD PER FI

Next, a new data set is created with one record per FI. This is done so we can print out the FI's ID in the tracking chart later on. All records are given a date value just before the beginning of the quarter so they will appear on the left margin of the tracking chart.

```
data fi;
      set roc (drop=code);
      length printval $11.;
      by fiid duid;
      if first.fiid;
      date=mdy((qtr*2)+(qtr-2),1,year)-2;
      after4=0;
      printval=fiid;
```

STEP 5 – CREATE ONE RECORD PER HOUSEHOLD

Then, a second new data set is created with one record per household. This is done so we can print out the household IDs in the tracking chart later on. All records are given a date value just before the beginning of the quarter so they will appear just to the right of the FI's ID from Step 4.

```
data du;
      set roc (drop=code);
      length printval $11.;
      by fiid duid;
      if first.duid;
```

```
      date=mdy((qtr*2)+(qtr-2),1,year)-1;
      after4=0;
      printval=duid;
```

STEP 6 – CREATE A UNIQUE ID FOR EACH RECORD

The first three sets are stacked together and a single ID is created based on the FI's ID, household ID, date, and whether the call occurred after 4:00pm. The variable PRINTVAL is created to equal the disposition code. The new dataset is sorted by the new ID and the date and time.

```
data new;
      set roc fi du;
      newid=compress(fiid||duid||date||after4);
      proc sort data=new;
            by newid date_time;
```

STEP 7 – USE LAG TO INCREMENT AN ID FOR EACH NEW INTERVIEWER/HOUSEHOLD COMBINATION

A variable named FIDU is created and incremented by 1 when a new NEWID appears, using the LAG function which compares the present record to the previous. A variable named CALL is created and set to 1 when a new NEWID appears. CALL is incremented by 1 while values of NEWID are the same, so all records with the same NEWID are ordered.

```
data new2;
      set new;
      if newid ne LAG(newid) then do;
            fidu+1;
      end;
      if newid ne LAG(newid) then do;
            call=1;
      end;
      if newid=LAG(newid) then do;
            call+1;
      end;
```

STEP 8 – CREATE AN ARRAY FOR ALL CALLS MADE BY AN INTERVIEWER TO A HOUSEHOLD DURING A HALF-DAY

An array is created with 9 variables relating to up to 9 calls made by a single FI to a single household during a single half-day period.

```
array codes(9) e1-e9;
      do i=1 to 9;
            codes(i)=0;
            if call=(i) then codes(i)=code;
      end;
```

STEP 9 – CREATE ONE RECORD PER INTERVIEWER PER HOUSEHOLD PER HALF-DAY

A new data set is created to put all calls by a single FI to a single household in a single half-day period in a single record.

```
proc means noprint;
      var e1-e9;
      by fidu;
      output out=new3 sum=e1-e9;
```

STEP 10 – CONCATENATE THE CALL DISPOSITIONS

The call records are put into a single string delimited by commas.

```
data new4;set new3;
      array codes(9) e1-e9;
      array outcomes(9) $ o1-o9;
      do i=1 to 9;
            if codes(i)=1 then outcomes(i)='NOT_HOME';
            if codes(i)=2 then outcomes(i)='UNAVAIL';
            if codes(i)=3 then outcomes(i)='REFUSE';
            if codes(i)=4 then outcomes(i)='COMPLETE';
      end;
```

```
      array outcomes2(8) $ o2-o9;
      array commas(8) $ c1-c8;
      do i=1 to 8;
            if outcomes2(i) ne '' then commas(i)=',';
      end;
      callstring=compress(o1||c1||o2||c2||o3||c3||o4||c4||o5||c5||o6||
            c6||o7||c7||o8||c8||o9,' ');
      keep fidu callstring;
```

STEP 11 – MERGE CALLSTRING DATA WITH DATE AND TIME AND DEDUPLICATE

The callstring data are merged back in with the date and time data. The records are deduplicated so only one record per FI per household remains.

```
data new5;
      merge new2 new4 (in=innew4);
            by fidu;
      if innew4;
      if callstring='' then callstring=printval;
      proc sort nodupkey data=new5;
            by fidu;
```

STEP 12 – CREATE X AND Y COORDINATE VARIABLES AND LABELS FOR THE EXCEL CHART

An ID is created for each FI and household combination, and the Y-coordinate variable is created based on it. The y-coordinate has .5 added to it so it will print between the gridlines, if used, not on them. The SAS date is converted to an Excel date by adding 21916 (Tilanus, 2004). For calls after 4:00pm, .5 is added to the date so the calls display in the right hand side of the cell for that day.

```
data new6;
      set new5;
      fiduid=compress(fiid||duid);
      if fiduid~=lag(fiduid) then do;
            v+1;
      end;
      y=v+.5;
      x=date+21916;
            if after4=1 then x=x+.5;
```

STEP 13 – EXPORT TO EXCEL USING ODS

The three output variables are exported to Excel using the ODS and PROC PRINT.

```
ODS HTML FILE='c:\temp\fichart.xls' RS=none STYLE=MINIMAL;
      proc print noobs data=new6;
            var x y callstring;
      run;
ODS HTML CLOSE;
```

The code in Steps 1 to 13 transforms the data into three variables suitable for charting:

- X - an x-coordinate variable equal to the value of the date of the call (ending in .0 when the call was before 4:00pm and .5 when after). When the record represents the beginning of a new set of records belonging to a new FI or household, DATE is set to a value just before the beginning of the field period of interest so the FIID and DUID will appear on the left hand side of the Excel chart.
- Y - a y-coordinate variable with a separate value for each FI/household combination in the data set so each record appears on a new line.
- CALLSTRING - a data label value equal to the FIID each time a new FI appears, DUID each time a new household appears, or the string of call dispositions for a particular FI and household at a particular day and time.

After the ODS exports the data to Excel, the data appear in the format depicted below.

X	y	Callstring
38167.0	1.5	999123
38168.0	1.5	IL99998817
38169.0	1.5	NOT_HOME
38171.0	1.5	REFUSE,REFUSE

38171.5	1.5	COMPLETE
38168.0	2.5	IL99998838
38169.0	2.5	NOT_HOME
38171.0	2.5	NOT_HOME
38173.0	2.5	NOT_HOME
38168.0	3.5	IL99999903
38169.0	3.5	NOT_HOME
38171.0	3.5	NOT_HOME,NOT_HOME
38173.0	3.5	NOT_HOME
38168.0	4.5	IL99999914
38170.0	4.5	NOT_HOME
38171.0	4.5	NOT_HOME
38174.0	4.5	COMPLETE

A pre-specified chart design populates the interviewer tracking chart, resulting in the chart shown below. This figure shows the work of two interviewers after one week of data collection. From the chart, a field supervisor can determine several things quickly about the behavior of these interviewers in the field. FI #999123 seems to display a rather typical pattern of behavior. Calls were made to several households on the same day. Many resulted in non-contacts but some resulted in completed interviews.

FI #999178 also made many calls before 4:00pm. This may indicate an inefficient work pattern. Also, it appears that at least one refusal was followed closely by a completed interviews. This is a very unlikely pattern, since those who refuse to be interviewed rarely change their minds later that day or the next day. Refusal conversion often requires a waiting period of at least a few days. The presence of such a pattern may suggest some protocol violation occurred. A supervisor reviewing this chart would likely recommend that this case and other similar ones for this interviewer be slated for follow up through a verification process to determine whether the interview was indeed valid.

Interviewer ID (FIID)	Dwelling Unit ID (DUID)	Thursday, July 01, 2004	Friday, July 02, 2004	Saturday, July 03, 2004	Sunday, July 04, 2004	Monday, July 05, 2004	Tuesday, July 06, 2004
999123	IL99998817	NOT_HOME		REFUSE,REFUSE COMPLETE			
	IL99998838	NOT_HOME		NOT_HOME		NOT_HOME	
	IL99999903	NOT_HOME		NOT_HOME,NOT_HOME		NOT_HOME	
	IL99999914		NOT_HOME	NOT_HOME			COMPLETE
	IL99999934	COMPLETE					
	IL99999945	NOT_HOME				NOT_HOME	
	IL99999955	UNAVAIL					
	IL99999965	NOT_HOME		NOT_HOME			NOT_HOME
	IL99999996	NOT_HOME				NOT_HOME	
999178	IL99998848	UNAVAIL		NOT_HOME		NOT_HOME	
	IL99998858	NOT_HOME UNAVAIL	COMPLETE				
	IL99998869	NOT_HOME	NOT_HOME,REFUSE,COMPLETE				
	IL99998879	COMPLETE					
	IL99998889	NOT_HOME,REFUSE		REFUSE			

CONCLUSION

This paper has shown that SAS software offers the capability to organize an enormous amount of survey process data into a simple and easy-to-interpret tracking chart for use by field supervisors and others in charge of reviewing survey data. The tracking chart makes it possible for one to detect patterns in interviewer behavior that may not have been realized with the data in their original format. The SAS LAG function and ODS were especially valuable in organizing the data.

We hope to further define this tool and investigate the possibility of adding color and embedded hyperlinks to the code displayed. This would empower the data reviewer even further and allow him or her to drill down into individual calls to read notes associated with them, or extract other important information.

As technology advances, the amount of process data that can be collected easily will continue to increase. In order to keep up with the possibilities these data represent, tools such as this one should continue to be developed. SAS software provides a means for developing such tools to ultimately ensure a high standard of data quality and efficiency in the survey process.

REFERENCES (HEADER 1)

Bosnjak, M. & Tuten, T.L. (2001). Classifying Response Behaviors in Web-Based Surveys. Journal of Computer Mediated Communication, 6.

Groves, Robert M., John Van Hoewyk, Grant Benson, Paul Schulz, M. Patricia Maher, Lynette Hoelter, William Mosher, Joyce Abma, and Anjani Chandra (2003) "Using Process Data from Computer Assisted Face to Face Surveys to Make Survey Management Decisions." Paper presented at the 2003 AAPOR Conference, Nashville, TN.

Heerwegh, D. (2004). Uses of Client Side Paradata in Web Surveys. Paper presented at the International symposium in honour of Paul Lazarsfeld (Brussels, Belgium June 4-5 2004)

Heerwegh, D. (2002). Describing response behavior in websurveys using client side paradata. Invited position paper presented at the International Workshop held by ZUMA in Mannheim (Germany), October 17-19, 2002.

Kalsbeek, William D., Steven L. Bottman, James T. Massey, and Pao-Wen Liu (1994) "Cost-Efficiency and the Number of Allowable Call Attempts in the National Health Interview Survey" Journal of Official Statistics, Vol. 10 (2), pp. 133-152.

Kulka, Richard A. and Michael F. Weeks (1988) "Toward the Development of Optimal Calling Protocols for Telephone Surveys." Journal of Official Statistics, Vol. 4 (4), pp. 319-332.

Lipps, Oliver, G. Benson (2005). "Interviewer Effort and Data Quality: Paradata Analysis of a Cross-National Survey: Working Paper 1_05" Draft paper downloaded on 8/31/05 from:
http://www.swisspanel.ch/file/working_papers/WP1_05.pdf

Purdon, Susan et. al. (1999) "Interviewers' Calling Strategies on Face-to-Face Interview Surveys" Journal of Official Statistics, Vol. 15 (2), pp. 199-216.

Safir, A., T. Black, and R. Steinbach (2001). "Using Paradata to Examine the Effects of Interviewer Characteristics on Survey Response and Data Quality." Proceedings of the Section on Survey Research Methods, American Statistical Association.

Scheuren, Fritz. (2001). "Macro and Micro Paradata for Survey Assessment" 1999 NSAF Collection of Papers Washington, D.C.: The Urban Institute. Assessing the New Federalism Methodology Report No. 7.

Tilanus, Erik. (2004). "Dating SAS and MS Excel." SUGI 29 Proceedings.
http://www2.sas.com/proceedings/sugi29/toc.html

Wang, Kevin, J. Murphy, R. Baxter, and J. Aldworth (2005). "Are Two Feet in the Door Better than One? Using Process Data to Examine Interviewer Effort and Nonresponse Bias" Forthcoming presentation at the Federal Committee on Statistical Methodology Research Conference.

Weeks, Michael F., Richard A. Kulka, and Stephanie A. Pierson (1987) "Optimal Call Scheduling for a Telephone Survey" Public Opinion Quarterly, Vol. 51, pp. 540-549.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Joe Murphy
RTI International

jmurphy@rti.org
www.rti.org