IMPROVING THE QUALITY OF SURVEY DATA: WESTAT'S COMPUTER ASSISTED CODING AND EDITING (CACE) SYSTEM

Dr. James E. Smith, Westat, Inc.
Dr. David L. Bayless, Westat, Inc.

Thousands of research surveys involving hundreds of thousands of respondents are conducted every year in the United States. The results of these surveys influence business strategies, public policy decisions, and in many other ways feed our information-hungry society.

But what can we say about the quality of these survey data? Recently, a senior technical official at the U.S. Bureau of the Census has said that "as the survey world currently exists, the concept of quality of survey data is not clearly defined and is certainly not measured well." [1] Thus, we are used to addressing questions about the quality of our automobiles, our workforce, and even our quality of life, but we are not so familiar with questions about the quality of our survey data.

DATA QUALITY

Of course everyone wants quality survey data. But, quality is one of those concepts, like happiness, whose exact meaning often escapes us. However, it is not so difficult for us to define lack of quality. We all have our anecdotes about lapses in quality -- like the ethnicity code that was inverted in the last half of a data file, leading to very interesting, but erroneous, results; or the large data file in which "<CR>" was typed dutifully into thousands of data fields by clerks who took their instruction manual too literally.

These are survey processing errors. They are examples of the kinds of mistakes that are inspected for, tested for, and hopefully found and eliminated through the checking of our survey data. The procedures for finding and fixing these types of errors are called, appropriately enough, "data cleaning."

Data cleaning is an inspection approach to data quality. It relies upon after-the-fact examination of the data to identify and correct errors. One of its main drawbacks is that all the costs associated with putting the errors into the data in the first place have already been incurred by the time data cleaning occurs. After-the-fact inspection does not reduce those costs. And, it does add the additional cost of trying to repair the data.

QUALITY CONTROL

Data cleaning is often done according to a quality control principles that can be called inspection methods. Tolerance limits are defined for the products and acceptance tests are conducted. If the product fails the test, remedial actions are taken to repair or re-do. This cycle is repeated until the product passes the acceptance test.

Modern methods of quality control, particularly the methods called statistical process control, have changed this approach. [2] They emphasize building quality into the data rather than inspecting errors out of the data. The objective of these methods is to measure and understand variation in the quality of the ontputs for the various processes that produce the product. Thus, the uew methods are based on continual analysis and monitoring of the production process in order to prevent the introduction of errors in the first place and improve data quality. There has been a shift toward quality improvement by continual process monitoring and a shift away from quality control by product inspection.

In order to apply the modern methods, fixed tolerance limits are replaced by "action limits" used to monitor the production process. Action limits define the range in which variation in some measured indicator of the production process is acceptable. If the indicator goes outside these limits, actions are taken to identify the source of this variation. Unlike fixed tolerance limits, the action limits are not used to gauge the *product* already made, but rather to monitor the ongoing *processes* that make the product.

In order to apply modern quality improvement methods it is necessary to understand the processes that produced the data. It is necessary to understand how the activities within the process may lead to quality or lack of quality in the product. And it is necessary to create and monitor indicators of the process in order to measure and control variation, especially when action limits are exceeded.

THE SURVEY DATA PRODUCTION PROCESS

As varied as surveys and survey methodologies are, there are enough commonalities in the ways in which most surveys produce data to make it feasible to lay out a general process. Most simply put, there are subjects "out there" from whom the researcher wants to "hear" certain data. The survey process involves "listening" to the subjects and "translating" the responses into a form useable by the researcher.

This translation process can be thought of as a very large, and usually costly, black box. (If the data do not emerge it is a black hole.) This black box usually involves several data collection, processing, and control processes. Traditionally these are listed as data collection, receipt control, coding and editing, data entry, and file building.

The survey process has come to rely heavily upon computers. Computer assisted data entry (CADE), computer assisted telephone interviewing (CATI), computer assisted personal interviewing (CAPI), computer assisted self administered questionnaires (CASAQ), and "disks by mail", are among the computerized survey methods that are in current use.

A general feature of these methods is that the data are placed into computer readable form early in the data collection process. Thus it has become feasible to enter the data more-or-less literally as received from the respondent and to let the computer assist in the coding and editing of the responses into the form required for analysis. Computer assisted coding and editing (CACE) is the name we have given to this process.

Depending upon the particular survey methodology, CACE may take place at various points in the process. For example, during a CATI survey, CACE may occur during the interview and provide immediate feedback to the interviewer about coding problems or data inconsistencies to be resolved with the respondent. In a mail survey, CACE may be performed during data entry as part of a CADE operation, or it may be performed after data entry as a separate batch or interactive process.

CACE AND SAS® SOFTWARE

Several systems for processing survey data have been developed with SAS® software. Some of the particular strengths of SAS® software have been its ability to handle character data and its powerful data step programming language for data manipulation. More recently the macro feature has significantly enhanced the power of SAS® software. However, a common feature of these systems is that they are batch oriented. This is not surprising, since SAS® software is itself essentially batch oriented despite a number of enhancements to make it more interactive.[3]

Our experience is that SAS® software is not suitable for implementing a highly interactive multi-user CACE system. However, SAS/QC® software is an efficient and sophisticated statistical process control tool for monitoring the CACE process.

Westat's CACE system, written in the C language, allows full interactive range-checking and consistency checking of survey data. SAS/QC® software is used to monitor the ongoing CACE process using the CACE journal file. This file is produced by CACE and records every edit operation, along with operator initials, time and date, and questionnaire ID.

ILLUSTRATIVE RESULTS

A CACE session begins with an operator (editing clerk) logging onto the system and identifying a file (batch of forms) to be edited. To assist in resolving errors the operator has the hardcopy forms for that file. Upon entering the ID of a form, the operator immediately sees a list of errors detected on the form. The examples we will present are from an actual large survey involving over 100 data items per form and about 60,000 forms. In this survey, nearly 400 edit checks were performed on each form. The operator usually waited one or two seconds from the time the form ID was entered until these checks were run and the list of errors appeared on the screen.

Edit checks and the errors they detect are usually divided into two categories: ranges and logics. Range edits check the values of individual data items to see that they are within acceptable bounds. Logic edits check for consistency of values between items. Both of these types of edits are performed in CACE, and operators see a listing with range errors indicated first, and logic errors indicated last. This encourages the operator to first resolve the usually simpler range errors before resolving logic errors.

The operator selects one error to resolve by moving the highlight bar over the selected error message. Upon pressing return, a new screen appears with the information necessary to resolve the error. For a range error this is the value of the variable and an indication of its legal range. For a logic error this is the values of each variable in the logic check. The operator consults the hardcopy survey form to see if the error is due to a data entry mistake, or some other obvious error that will be resolved by changing the value in the file to match the value indicated by the respondent on the form.

If the operator can see no such mistake that accounts for the error, then an "override" condition exists. This means that the data in the file are a faithful representation of what is on the respondent's survey form. In this case the operator must specifies an override for the range or logic error. An override is a flag that will suppress the error message and can be used to identify the need for further editing of the data at a later stage.

Thus, the job of the CACE operator is to examine each error reported by the system for each form, and then to either resolve the errors by changing erroneous data values to correspond with the hardcopy survey form, or to override the errors in order to indicate that there is no data entry error and the errors must be handled by some later editing process within CACE (which we will not describe further here.)

In order to monitor this editing process, SAS/QC® software is used to produce control charts from the CACE journal files. Figure 1 shows a control chart for the number and standard deviation of range changes (corrections or overrides) made by CACE operators. The purpose of this chart is to monitor the numbers of range checks that operators are dealing with, and to determine if these numbers are consistent across operators. It can be seen that there are points on this chart that indicate the process does vary outside of control limits for several operators. This is not an operator problem *per se*, but it is a management problem, since the burden of performing range checks seems to be falling on some operators more than others.

Figures 2 and 3 take the process one step prior to the CACE operation. These figures show the number of range errors and logic errors detected for the field data collectors who submitted the forms. These charts are useful for signalling problems with particular field data collectors, or as is the case in these charts, particular problems with the types of respondents different interviewers are contacting. In this survey, interviewers were obtaining the data from postsecondary schools, and certain interviewers went to certain types of schools, some of which had more accurate data than others.

One of the risks of a CACE process is that operators will become "hypnotized" by the routineness of the process, and will start to form a pattern of responses to the computer that is not properly resolving errors. For example, the operator could respond to range error messages by immediately overriding the error rather than first carefully checking the hardcopy form. Figure 4 is a control chart designed to detect this. Certain operators are outside the upper control limits for the number of range errors resolved by overriding the error message. This signals the need to examine the work of these operators to determine the probable causes of this condition.

Another operator behavior pattern that deserves monitoring is the possible effects of fatigue on data editing. This is indicated in Figure 5 where number of range overrides is monitored by time of day. The reasoning behind this control chart is that operators may tend to immediately override more errors without checking the hardcopy when they are more fatigued. The control chart does in fact show a significant increase in the number of range overrides in the late afternoon. This signals the need

for further identification of causes. When this was done, it was found that two factors were operating: operators did apparently immediately override more errors in the late afternoon *and* particularly messy batches of data tended to be saved until the afternoon thus generating more overrides due to the nature of the data.

Statistical process monitoring and control using SAS/QC® software has proven highly informative in the editing of survey data with the aim of improving the process, and therefore the ultimate quality of the data. Unlike older methods of inspection that focused on data cleaning as the objective, newer methods focus on continual monitoring and control of the process of survey data editing. The result is more involvement and better performance from data editors, and greater understanding of the factors that determine survey data quality.

FOOTNOTES

[1] Bailar, Barbara A., "The Quality of Survey Data" *1984 Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pp. 43-52.

[2] Shewhart, Walter A., *Statistical Method from the Viewpoint of Quality Control*, New York, Dover Publications, 1986.

[3] Blank, Grant, "The Obsolescence of the SAS System" Paper Presented at the SAS® Users Group International, Tenth Annual Conference (1985), pp. 285.

Figure 1

CACE OPERATOR CONTROL CHART: Range Changes

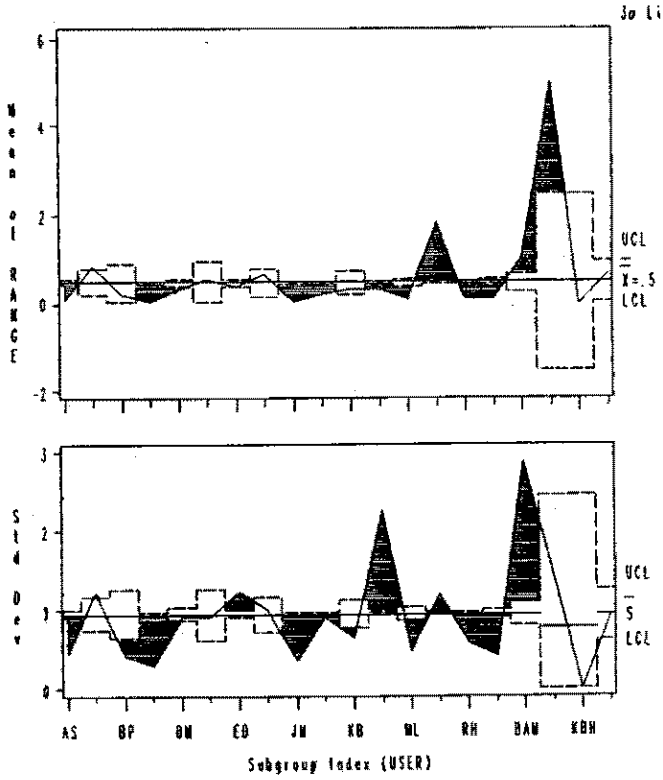

Figure 2

FIELD STAFF CONTROL CHART: Range Changes

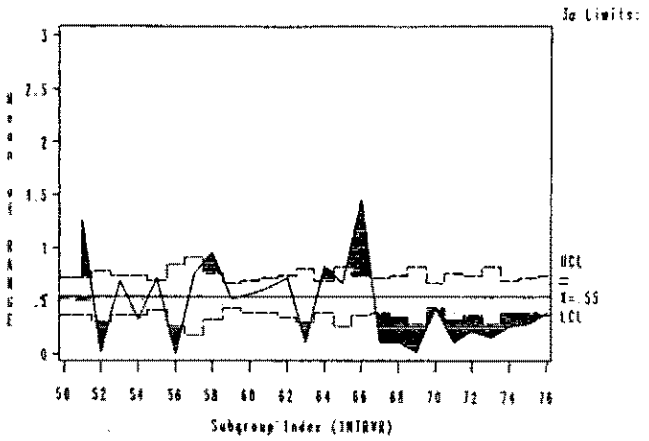

Figure 3

FIELD STAFF CONTROL CHART: Logic Changes

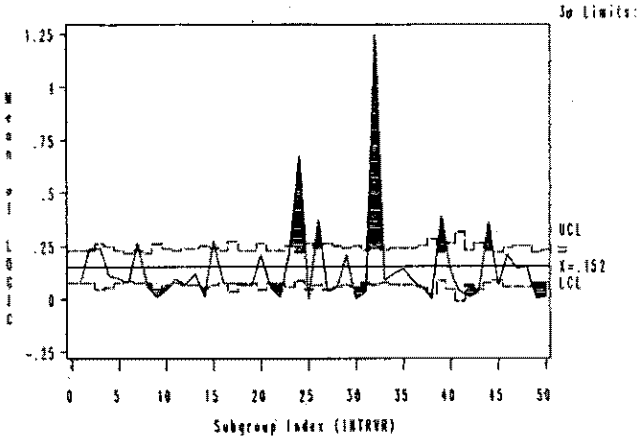

Figure 4

CACE OPERATOR CONTROL CHART: Range Overrides

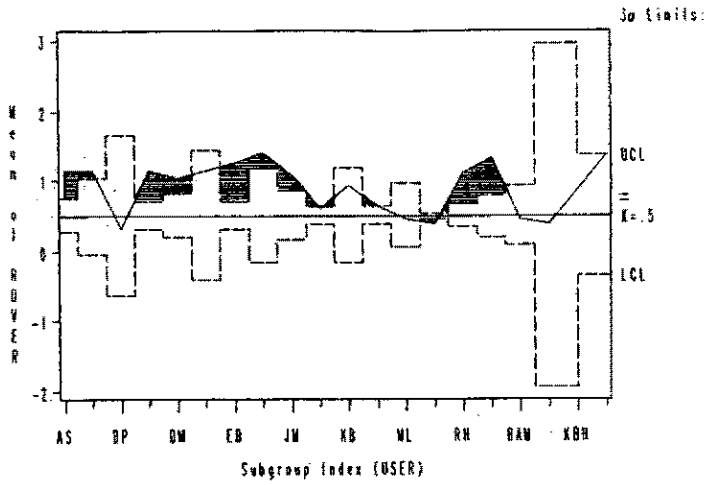

Figure 5

HOUR OF DAY CONTROL CHART: Range Overrides