United States Department of Agriculture

National Agricultural Statistics Service

Research Division

SRB Research Report Number SRB-95-05

August 1995

The Effects of Editing and a Review of Warning Errors From a CAPI/IE Perspective

Tom Pordugal

THE EFFECTS OF EDITING AND A REVIEW OF WARNING ERRORS FROM A CAPI/IE PERSPECTIVE, by Tom Pordugal, Ohio Applications Research Section, Survey Research Branch, Research Division, National Agricultural Statistics Service, United States Department of Agriculture, Washington, D.C. 20250-2000, July 1995, Report Number SRB-95-05.

ABSTRACT

In 1994, the Technology Research Section of the National Agricultural Statistics Service (NASS) directed a study involving all June area segments in Indiana. The research focused on evaluating Computer Assisted Personal Interviewing (CAPI) as a data collection method as compared to the traditional Paper and Pencil Interviewing (PAPI) method. A test was conducted in Indiana that examined the collection of data from a sub-sample of their area segments through CAPI. The balance of area segments ran through a Blaise interactive edit (IE) after one statistical/clerical review of the paper questionnaires. All data were edited interactively by survey statisticians prior to the mainframe Survey Processing System (SPS) edit. This paper addressed two post-survey issues. The first issue involved the differences in the way that non-critical errors are handled between interactive editing and batch editing systems (SPS). Large amounts of warning errors cannot be ignored by the interactive editing system, and the analyses in this paper identified such warning errors. The second issue addressed data quality within the CAPI and PAPI data and differences between CAPI and PAPI data collection techniques. The analyses compared the CAPI and PAPI data sets to the respective data in the clean and edited data set.

KEY WORDS

Computer Assisted Personal Interview (CAPI); Interactive Editing (IE); Binomial Proportion Test; Sign Test.

ACKNOWLEDGEMENTS

The author thanks Roberta Pense and Mark Schleusener for their assistance with analysis and helpful suggestions. The author also thanks Guy Pense for his assistance with data processing.

TABLE OF CONTENTS

INTRODUCTION

In 1994, the Technology Research Section conducted a study to explore two methods of improving the quality of survey data. These involved the use of Interactive Editing (IE) and Computer Assisted Personal Interviewing (CAPI). Although the Technology Research Section conducted the June Area Survey research project in Indiana and Pennsylvania, this paper's study was limited to only Indiana's June Area Survey. The purpose of the study in Indiana was to pilot test CAPI and IE procedures on a small scale. This paper deals with two post-survey issues. The first issue focuses on the differences in the way that non-critical errors are handled in interactive editing and NASS's batch editing systems (SPS). The second issue addresses data quality within the CAPI and PAPI data and the differences between CAPI and PAPI data collection methods.

Granquist (1995) defines editing as the procedure for detecting and adjusting individual errors in data records resulting from data collection and capture. The checks for identifying missing, erroneous, or suspicious values in computer-assisted editing are called edits. Moreover, editing change refers to the situation where an item value (a question from the survey) is adjusted as a consequence of action taken when an error is identified.

NASS's traditional method of collecting and editing data for face-to-face interviews is through paper-and-pencil interviewing (PAPI) and editing with Survey Processing System (SPS) batch processing on a mainframe. Generally, editing survey data involves looking at whether the SPS generates critical or non-critical errors. Non-critical errors are called warning errors (or messages) or soft errors, where the data are considered acceptable, but unusual, by the SPS. Critical errors (or hard errors) occur in the data when the upper and lower bounds set by the system are exceeded or when the data are not recognized by the system. Non-critical errors disappear after the first SPS edit unless there is an update to the record.

With batch processing, the errors detected after data entry involve many steps: rehandling the questionnaire; writing the corrections on paper; sending the corrections back to data entry; and running another SPS edit usually overnight. Therefore, the State Statistical Offices spent much time reviewing the questionnaires before data entry to avoid corrections after data entry.

The interactive process edits and displays data on a computer screen dynamically. As all data involved in the error are displayed on the screen, the enumerator or data editor can choose which data values to correct. As the statistician or statistical assistant makes corrections and/or changes to data, the record is immediately re-edited. This gives the editor immediate feedback. Calculations of ratios can be carried out in error messages, and recalculated at will. Derived values such as yields can become part of the display as can historical or list frame data. Unlike the SPS, the interactive editing process eliminates re-sending forms to data entry or waiting for overnight edit runs.

CAPI refers to the data collection method. Instead of using paper and pencil to record the information, enumerators use sub-notebook computers loaded with Blaise programs that contain the questions, routes, and edit checks. Because there are editing checks during the interview, interactive editing does occur within CAPI. However, after the data are sent to the State Office, more stringent edits can also be applied interactively using the same software. For purposes of this analysis, this type of editing in the State Office is defined as interactive editing.

The seven supervisory enumerators in Indiana chose one enumerator from each group of enumerators to do CAPI. Enumerators were selected by their supervisor to participate in the project, and two enumerators had used CAPI in 1993. Although the enumerators' computer backgrounds and aptitudes ranged widely, they were not just chosen at random.

Once chosen, the enumerators were to complete their whole assignment by means of CAPI. All other enumerators used paper and pencil to collect the data. The seven enumerators completed a total of 363 interviews with CAPI. The interviews classified as non-agricultural were screen-out interviews. The agricultural interviews were those where actual farm data were collected. Twenty-three percent of all agricultural interviews were collected by CAPI.

WARNING ERRORS AND THE EDITING POLICY

Purpose

The purpose of this analysis was to identify and examine the warning-type errors that were generated in both the unedited PAPI data and the PAPI portion of the edited file. This would indicate unnecessary data warnings, i.e., where the data are not being changed very often. The unedited PAPI records were targeted because they were the "dirty" records, i.e., those records that passed through the cursory pre-keyentry review. Thus, the edits that rarely resulted in changing the data might be eliminated entirely in the edit by letting the analysis packages or data listings identify the problems. Moreover, based on the frequency of edit checks, the edit checks can be modified so they do not result in a data review as often. In addition, changes could be made to the survey question to avoid recording erroneous data in the first place.

It is always important to limit error review to those data values most likely to be in need of correction. Reviewing a large number of error messages is tedious. Repetitive and monotonous actions tend to have editing "mistakes" creep in. Editors' time should be spent efficiently, usually looking at "incorrect" data. This is true of any editing system.

With NASS moving to interactive editing, producing efficient edits is even more important because interactive editing handles errors differently. Error suppression in the Blaise interactive edit takes some work not required in batch editing. If the statistician "does nothing" in SPS, then the warning goes away, i.e., the data are accepted with no corrections. On the other hand, in interactive editing the statistician has to hit the "s" key to indicate the warning has been reviewed and the data accepted. If there are a large number of unnecessary warning errors, this can become time consuming. However, this ensures that warning errors are reviewed.

Methods

This part of the research looked only at paper questionnaires (PAPI). The CAPI data were not examined because they already had been through editing checks. With CAPI, edits occurred during each interview, so there would be no way of knowing how many error corrections were made during each interview.

A minimal amount of editing was done on the paper questionnaires before the "unedited" data files were processed through the computer edit. This editing considered the following: checking and verifying the overlap/non-overlap coding on the questionnaires; filling a cell if the enumerator checked the box, but forgot to code the cell; marking the appropriate completion code boxes; entering "-1's" where the enumerator wrote "don't know"; making sure the crops are recorded in the proper cells; entering item codes for optional land uses codes in the Field Crops table; reviewing for legible entries; and using historic data or calling the respondent to obtain and/or verify necessary data.

The data as originally keypunched were processed through the SPS edit. Corrections to administrative-type data, such as strata codes, were processed as well. This made sure that the commodity data were routed through the edit correctly. Namely, this ensured that the data would be present when required. However, any adjustments to the original commodity data were not processed.

The counts of non-critical SPS errors in the unedited data were compared to the errors in the edited PAPI data to see roughly how often the warnings resulted in changed data. However, the unedited data may have caused a critical error in the first edit, and the correction may have generated a warning error in the edited file edit. Thus, it would be erroneous to say the following: of the "x" warning errors in the unedited file, "y" are still present in the final data set.

Counts were identified for those errors that occurred at least nine times in either the unedited or edited data and did not appear to result in changed data at least 50 percent of the time. If the statistician needed to change the data, then the interactive edit was much easier. However, if the data were not changed, then the interactive edit required more work on the statistician's part to suppress them. This was the reason that "at least 50 percent of the warning errors" was targeted, i.e., that the data generally were changed at least 50 percent of the time.

Analysis

Table 1 shows the counts of the non-critical SPS errors of the PAPI data. Counts were identified for those errors that occurred at least nine times in either the unedited or edited data and did not appear to be corrected at least 50 percent of the time. However, any further analysis can only be taken lightly, since the edited data does not necessarily show how many records actually got changed.

Table 1: Counts of Non-Critical Errors for Unedited and Edited Data for PAPI

Error Counts	Unedited	Edited
Winter Wheat Acres Planted for Grain: Warning 249	17	9
Pigs on Hand March Thru May: Warning 511	33	21
Calves Born Since January 1 (Farm): Warning 605	16	8
Cattle on Feed Total (Farm): Warning 616	35	16
Rams and Wethers 1 Year and Older: Warning 709	9	9
Ag Land Values (IC416): Warning 864	317	284

Discussion and Recommendations

Based on the research, agricultural land values produced the most warning errors. Based on Table 1, this warning was the only "bad" one, i.e., where corrections were hardly ever made to the data. However, this was the first year that the Ag Land Values Section was added to the June Survey.

Thus, the **first recommendation**, informally made to the Statistical Methods Branch in early 1995, was to reconsider the edit checks for the warnings listed in Table 1. Steps had already been taken based on survey evaluations to modify errors 864 (ag land values) and 616 (cattle on feed). The other warning errors were considered prior to the 1995 edit.

The **second recommendation** involves pre-testing the editing instrument. Agricultural land values section was new to the June Area Survey in 1994. The questionnaire was pre-tested, but the edits were not. Generally, the focus of edit instrument pre-testing should be the following: how often certain warnings types surface, i.e., how often warnings need to be corrected; the clarity of error messages; and whether, in a data collection environment, an enumerator can get a critical error in a situation that is possible.

Pre-testing editing instruments and data collection instruments that contain edits and instructions are particularly important if they are used in a distributed environment. If changes get made to the instrument, then the instrument has to be re-distributed, which means re-installing in the State

Office. This, in turn, requires more work for the State Office.

The **third recommendation** involves the building of a tracking mechanism into the edit instrument to track whether edits are effective. The tracking system might consist of counts of the errors, warnings corrected, and warnings suppressed, or the tracking system might involve automatically capturing the unedited data and comparing it to the edited data. Using current procedures, Headquarters will not get counts from the SPS "unedited data errors" once interactive editing is fully in place.

EFFECTS OF EDITING

Purpose

The purpose of this part of the research dealt with the issue of data quality and differences between the CAPI and PAPI data collection procedures. If potential data problems can be resolved at the interview level with the farm operator or other respondent, then less clerical and/or statistician editing will be necessary. The following analysis used CAPI data as it was received from the interviewers, i.e., the data gathered at the interview level with the respondent where some interactive editing occurred.

Methods and Analysis

This issue addressed data quality via the use of CAPI. The June area tracts were split into two datasets: CAPI and PAPI. The analysis only considered positive data by item of interest; refusal or inaccessible records were omitted. The variables listed in the following table (Table 2) were the items of interest in the research.

Since the area segments were not randomly assigned to the enumerators, the only feasible way to look at data quality was to use "before" and "after" data, i.e., to look at the amount of editing going on in the office. The CAPI and PAPI data after IE comprised the "after" data. The "before" data for PAPI consisted of the unedited data from the agricultural tracts. The "before" data for CAPI consisted of the unedited CAPI data, before the office interactive edit.

For each item of interest, difference variables were created. The difference variable was calculated as the "value after office editing" minus the "value before office editing". To examine data quality, the research looked at "after minus before" data for CAPI and PAPI separately to see how many changes were made in editing for each data collection method. The assumption was that "after" editing was the truth. By creating these difference variables for CAPI and PAPI, the number of matches (no editing changes) and editing changes (positive or negative) were generated. In this way the research addressed the issue of data quality. Namely, if a great deal of editing was required, then the data collection quality was poor.

By creating these difference variables for CAPI and PAPI, nonparametric statistical tests could be constructed. The best way to look at level differences in the amount of editing between CAPI and PAPI data was using binomial proportional tests, assuming

Table 2: Variables (Items of Interest) From the June Area Survey

Agricultural Tracts Only	Agricultural Tracts Non-Overlap for Hogs/Crops	Agricultural Tracts Non-Overlap for Cattle/Sheep
Total Tract Acres (IC840)	Total Grain Stocks (IC121+IC125+IC126)	Total Beef Cows (IC351)
Total Farm Acres (IC900)	Total Tract Hogs (IC200)	Total Milk Cows (IC352)
Total Waste Acres (IC841)	Total Farm Hogs (IC300)	Total Cattle/Calves (IC350)
Winter Wheat Planted Acres (IC540)	Farrowing Sows (IC326)	Cows Expected to Calve (IC361)
Winter Wheat Acres Planted for Grain (IC541)	Pigs on Hand From Litters (IC327)	Total Calf Crop (IC362)
Oats Planted Acres (IC533)	Pigs Slaughtered or Moved Off Land (IC328)	Total Cattle on Feed (IC652)
Oats Acres Planted for Grain (IC534)	Slaughter Pigs (IC320)	Total Cattle/Calves in Field (IC250)
Corn Planted Acres (IC530)	Feeder Pigs Purchased (IC340)	Total Sheep on Feed (IC698)
Corn Planted Acres for Grain (IC531)	Total Dead Weaned Pigs (IC355)	Total Stock Sheep (IC043)
Soybean Planted Acres (IC600)		Total Ewes to Lamb (IC289)
Soybeans Following Another Crop (IC602)		Total Lambs Born (IC288)
Alfalfa Hay Acres (IC653)		
Grain Hay Acres (IC656)		
Other Hay Acres (IC654)		

large populations and different variance populations. Proportions were constructed for each item of interest (Table 2) based on the number of CAPI and PAPI editing changes. Binomial proportions tests were chosen because the ratio of agricultural tract interviews collected through PAPI to those collected through CAPI was roughly three to one.

The binomial proportions test showed if there was any significance in the proportional differences in the amount of editing between the CAPI and PAPI data. For each item of interest, data quality referred to the proportions of edit changes in the CAPI and PAPI data. The hypothesis tested was the following: CAPI and PAPI approaches to data collection produce different quality data, based on proportions. Namely, the test was whether the same amount of editing (proportionally) was occurring in the CAPI versus the PAPI mode of data collection. The procedure is thoroughly described in Appendix A.

The results of the binomial proportions tests showed the following: Thirty-four variables were tested for significance (from Table 2), and of those variables, none showed significance. Because of the large sample sizes, Z-statistics were constructed at the $\alpha = 0.05$. The total milk cows variable (IC352) showed significance but had a small sample population ($n=2$) in the CAPI portion. Thus, the analyses showed that there were no significant differences in the proportional differences, i.e., no significant differences in the proportion of editing changes of the CAPI and PAPI data. Thus, for each of the thirty-four variables, the same amount of editing was required for both the CAPI and PAPI data collection modes.

Given that the same amount of editing is occurring in the CAPI and PAPI mode of data collection, the research addressed whether the editing changes were mostly positive, negative, or both. Specifically, the research focused on whether the statistician or data editor changed the data randomly about zero. The sign test for medians was used to test whether the median of the data was significantly different from zero for each item of interest and data collection method (CAPI/PAPI). The sign test does not assume symmetry about a median.

Distribution-free methods were used in the research because the distribution of the differences variables were not known. According to Moussa (1991), if the underlying distribution is unknown, or when normal theory assumptions are not met, nonparametric methods are more efficient and powerful than their normal-theory counterparts. Moreover, if the true distribution is normal and one uses nonparametric methods, the loss in efficiency and/or power is usually not substantial. The research chose nonparametric methods for these reasons. The sign test procedure is described in Appendix B.

For each variable the editing changes were tested, using the sign test, based upon the number of positive or negative departures from zero. The tests were done separately for CAPI and PAPI. Significance was found when the editing changes deviated more positively than negatively or vice versa. In other words, significance implied that the statistician was making some consistent change.

Table 3 shows the counts and nonparametric statistics of editing changes for those variables which showed significance for either the CAPI or PAPI mode of data collection. None of the CAPI variables were significantly different from zero, but two of the PAPI variables, namely, total farm acres and soybeans following another crop, showed significance at the $\alpha = 0.010$ level. As shown in the table, the editing changes were both in the positive and negative direction but not necessarily symmetric with respect to zero.

Table 4 shows the counts of the magnitude and direction of the PAPI editing changes in Table 3. The CAPI data are also shown. For PAPI, the total farm acres variable (IC900) was decreased by at least 100 acres by the statistician in five instances. Similarly, the statistician increased the acreage seven times. Thus, "outliers" (at the "at least 100 acres" interval) were evenly distributed. Small changes, i.e., changes less than or equal to ten acres, were almost always positive. The statistician may have increased the farm acreage slightly when the farm operator lived outside the segment, but the original data indicated the tract and farm acres were the same.

Table 4 also shows no changes in the CAPI data for soybean acres following another crop (IC602), while the PAPI data are reduced by the statistician on every occasion. There may have been reporting errors with double-cropped soybeans in the PAPI mode of data collection. The editing changes showed that these acres decreased by at most double to that of soybean planted acres (IC600). The enumerator probably was adding the double cropped acreage (from IC844) to this item code.

Discussion

CAPI and PAPI data quality were tested by the number of times the statistician increased or decreased the value. Based on the number of editing changes, both the CAPI and PAPI data collection methods proved to be reliable data collection tools.

This research supported CAPI data quality for the 1994 June Agricultural Survey in Indiana. Since few editing changes were made to the CAPI data, the analyses showed that most of the potential data errors were eliminated at the face-to-face interview level with the respondent.

Table 3: Counts of Editing Changes and Nonparametric Statistics for Those Variables Which Showed Significance

Counts p-Value	CAPI Editing Changes				PAPI Editing Changes			
	Total	Positive	Negative	None	Total	Positive	Negative	None
Total Farm Acres: IC900	6 0.2188	5	1	316	32 0.0070*	24	8	1070
Soybeans Following Another Crop: IC602	0 -----	0	0	8	16 0.0001*	0	16	32

* indicates significance at $\alpha = 0.010$

Table 4: Counts by Interval of the Magnitude and Direction of the CAPI/ PAPI Editing Changes in Table 3

Counts by Interval	CAPI IC900	PAPI IC900	CAPI IC602	PAPI IC602
Acres < -1000	0	3	0	0
-500 >Acres>= -1000	0	0	0	0
-100 >Acres>= -500	0	2	0	0
-50 >Acres>= -100	0	2	0	0
-10 >Acres>= -50	0	1	0	14
0 >Acres>= -10	1	0	0	2
0 <Acres<= 10	3	13	0	0
10 <Acres<= 50	0	4	0	0
50 <Acres<= 100	0	0	0	0
100 <Acres<= 500	0	5	0	0
500 <Acres<= 1000	2	2	0	0

REFERENCES

Ferguson, Dania; Graham, Mitch; Hohenbrink, Rita; Matthews, Wayne; Pierzchala, Mark (organizer); Schuchardt, Rick; Stewart, Roger; and Yost, Mickey (1990), Interactive Editing Research, A Report and Recommendations by the Interactive Editing Working Group, United States Department of Agriculture, National Agricultural Statistics Service.

Granquist, Leopold (1995), Chapter 21 of Business Survey Methods, Improving the Traditional Editing Process, John Wiley & Sons, New York.

Hollander, Myles, and Wolfe, Douglas A. (1973), Nonparametric Statistical Methods, John Wiley & Sons, New York.

Moussa, Effat A. (1991), Computer Integrated Probability and Statistical Methods for Science and Management: Part II, DePaul University, Chicago.

Mergerson, Jim; Pense, Roberta; Pierzchala, Mark; Romeo, Ann; and Schleusener, Mark (1994), "The 1994 June Area Frame Project Experience: Interactive Editing and Computer Assisted Personal Interviewing" (unpublished technical report), United States Department of Agriculture, National Agricultural Statistics Service.

Federal Committee on Statistical Methodology and Subcommittee on Data Editing in Federal Statistical Agencies (1990), Data Editing in Federal Statistical Agencies, Statistical Policy Working Paper 18, Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget.

SAS Institute Inc. (1988), SAS\STAT User's Guide, Release 6.03 Edition.

SAS Institute Inc. (1985), SAS User's Guide: Basics, Version 5 Edition.

APPENDIX A: THE BINOMIAL PROPORTIONS TEST

Suppose that comparisons are made on the proportion of CAPI and PAPI editing changes. Namely, a random sample of n_1 items is drawn from the CAPI population and Y_1 of them are editing changes. Another random sample of n_2 items is drawn from the PAPI population and Y_2 of them are editing changes. If the assumption is made that the true proportions of the CAPI and PAPI editing changes are p_1 and p_2, respectively, then Y_1 and Y_2 are binomial distributions with parameters n_1 and p_1 and n_2 and p_2, respectively.

Hypotheses tests are constructed about the values of p_1 and p_2 , where p_1 and p_2 are two independent proportions of CAPI and PAPI, respectively. Then the null hypothesis, H_o: $p_1=p_2$, is tested against the alternative hypothesis, H_1: $p_1 \neq p_2$, at α-level of significance. The best estimators of p_1 and p_2 are $p_i' = Y_i / n_i$, for i=1 and i=2, and each estimator is approximately normal with parameters p_i and $p_i q_i / n_i$. The difference, $p_1'-p_2'$, has an approximate normal distribution with parameters (p_1-p_2) and $(p_1q_1/n_1 + p_2q_2/n_2)$. Another way of writing the null hypothesis is the following: H_o: $p_1-p_2=0$. The variance of $(p_1'-p_2')$ is $pq/n_1 + pq/n_2$. To estimate p and q under H_o, the two samples are pooled and the test statistic reduces to the following:

$$Z_o = \frac{(p_1' - p_2')}{\sqrt{\frac{p'q'}{n_1} + \frac{p'q'}{n_2}}}$$

and the decision rule is to reject H_o if $Z_o > z_{1-\alpha/2}$ or $Z_o < z_{\alpha/2}$. The pooled estimate becomes $p' = (Y_1 + Y_2)/(n_1 + n_2)$.

APPENDIX B: THE SIGN TEST FOR MEDIANS

Given a random sample, $X_1, X_2, \ldots, X_n$, from a distribution, $F_X(x)$, which is unknown but is thought to be continuous and dependent on a location parameter. In the research, the hypothesis tests that the median of the distribution is $M=0$. Namely, the test is the following: Test the null hypothesis, $H_o: M=M_o=0$, against the alternative hypothesis, $H_1: M \neq M_o \neq 0$, and let $D_i = X_i - M_o = X_i$, where $i = 1,2,\ldots n$. Under H_0, the distribution of X has median M_o, and hence the differences, D_i, have median zero. In other word, if H_o is true, then the signs of the differences will be equally likely to be positive as negative. It is expected that the number of positive differences should be close to the number of negative ones.

If Y_+ and Y_- represent the positive and negative differences (positive editing changes and negative editing changes) in the sample for each item of interest, respectively, and $p = P[X_i \geq M_o = 0]$, then under H_o, the distribution of the statistic Y_+ (or Y_-) is binomial with parameters n and ½. In this research the test statistic is the smaller of the Y_+'s or Y_-'s (or $\min(Y_+, Y_-)$) because there is a two-sided alternative hypothesis. The decision rule is then to reject H_o if the p-value of the test statistic is too small, i.e., $p < 0.05$. If the sample size n is small, i.e., the number of positive and negative editing changes is less than 10, then the binomial distribution can be used to evaluate the exact p-value of the test statistic. For the binomial distribution, the decision rule is to reject H_o if:

$$P(\text{p-value}) = 2\sum_{y=0}^{y_o} \binom{n}{y} \left(\frac{1}{2}\right)^n \text{ is too small,}$$

where $y_o = \min(Y_+, Y_-)$. If the sample size is large, i.e., $n > 10$, then the normal distribution can be used to approximate the null distribution of the sign test statistic. For the normal distribution, the decision rule is to reject H_o if the test statistic:

$$Z_o = \frac{Y_o - \frac{n}{2}}{\sqrt{\frac{n}{4}}} \text{ is smaller than } Z_\alpha$$

for $|Z_o| > Z_{1-\alpha/2}$.