CONVERSION OF A MAJOR NASS PROBABILITY SURVEY TO BLAISE

Asa Manning
National Agricultural Statistics Service, USA

1. Organizational issues

The National Agricultural Statistics Service (NASS) is an agency of the United States Department of Agriculture. NASS is responsible for collecting data and making estimates relating to the nation's agriculture. The agency has staff located in Washington, DC and in offices around the country.

NASS has a field office in almost every state, called a State Statistical Office (SSO). The office in New Hampshire services six of the small New England states out of one location, thus there are a total of 45 SSOs. Each office is responsible for collecting data within their state(s) and setting recommendations for commodities grown, raised or produced there. Once the SSOs complete their state(s) recommendations, they are then transmitted electronically or mailed to headquarters (HQ) in Washington, DC. In HQ a group of statisticians, representing HQ and the SSOs, set a national level estimate. The state recommendations are then reviewed and revised as needed to sum to the national level. Publications, containing the final estimates at the state and national levels, are then released to the public.

This distributed structure presents some challenges for HQ staff responsible for coordinating the surveys at a national level. An example is the Computer Assisted Survey Section (CASS) which is responsible for the development and maintenance of the instruments used for Computer Assisted Telephone Interviewing (CATI) and Interactive Editing (IE) for national survey applications. The diverse needs of the SSOs create some problems that must be dealt with.

First, the questionnaires differ from SSO to SSO. In effect CASS must write forty-five instruments instead of one. NASS is trying to use an ap-

proach described in a paper presented at the 1992 International Blaise Users Conference by Mark Pierzchala. The concept in simple terms involves developing a library of tested code. A CAPI instrument is then used to collect specifications for each state. The specifications are then used by a questionnaire generator to assemble the final instrument for each state in an automated fashion using the library of code. This procedure holds tremendous promise for certain NASS applications.

Another effort which will reduce coding is the development of a set of standard shells. NASS has a standard approach to collecting administrative data on its surveys. CASS is in the process of developing code to handle these questions, as well as setting up survey management data in a standard way. When completed, these shells will be used for almost every Blaise application within NASS. A detailed paper on this effort, titled "Standard Multi-Survey Shells in NASS", will be presented at this conference by Mark Pierzchala and Roger Schou.

Each SSO will also have its own idea of the best way to manage the survey in their office. Some offices will mail questionnaires while others will not. CATI will be used on a large portion of the sample in one office, but much less in another. Editing may be the responsibility of a few staff or many. These differences mean that any plan that CASS designs must be very flexible, in many cases providing more than one solution for a single issue.

2. History of Computer Assisted Survey Information Collection (CASIC) in NASS

NASS was one of the original partners with The University of California at Berkeley in the development of the CASES software. This relationship started in the early 1980's. NASS now uses CASES for its CATI applications in all SSO's.

NASS began research with Blaise in 1987. The research has been concerned with interactive editing and CAPI. Blaise has been adopted as an interactive

editing tool by NASS. In 1992 NASS management decided that Blaise should be evaluated as a possible replacement for CASES, since it could do both data collection and editing.

During the last several months, CASS has planned a full scale test of Blaise on its September Agricultural Survey. This test will hopefully resolve once and for all the question of Blaise versus CASES.

3. The application

The cornerstone of NASS's estimation program is the Agricultural Survey. This survey is conducted quarterly (March, June, September and December) in all states, except Alaska and Hawaii. Information is collected on crop acreage and production, grain stocks stored on farms, and hogs. Related livestock surveys are conducted in January and July for cattle, sheep and goats.

This application itself presents some complex issues that CASS has to deal with. These surveys involve multiple sampling frames, extremely tight schedules, multiple modes of data collection and large sample sizes. Each of these issues will now be looked at in more detail.

The sample comes from two frames, list and area. A major portion of the sample is selected from a stratified *list* of farm operations that each SSO maintains for their state(s). During the June Agricultural Survey, all farmers operating tracts of land within sampled *areas* of land, called segments, are interviewed. If an operation is found in one of the segments that is "not on the list", then that operation is classified as NOL. The NOL tracts provide a measure of the incompleteness of the list, and are then included in the sample for each of the follow-on surveys (September, December and March). The multiple frame design adds several administrative type questions needed to detect possible overlap between frames. Extensive coding in the multiple frame shell was required to handle these questions.

The Agricultural Survey operates under an extremely tight time schedule. The Reference Date for each survey is the first day of the month. From the start of data collection to the publication of the first estimates is about four weeks. The time allowed for data collection and editing is only the first 14-18 days of the month. This restricted schedule means that the SSOs must use every hour of time wisely.

As a result of the brief data collection window, the SSOs use a combination of data collection modes:

Mail – If a sufficient response rate can be achieved, questionnaires will be mailed. Respondents may return 20-30 percent of the mailed questionnaires. Use of mail will vary from office to office.

CATI – On the evening prior to the first day of the month, the SSO will begin CATI from the office. The SSO will usually designate certain samples to not attempt in CATI and sends these questionnaires out to the staff of field interviewers. CATI will continue for 4-8 days. At the conclusion of CATI all remaining samples that have not been returned by mail or completed using CATI are turned over to the field interviewers.

PAPI – Paper questionnaires will be used for all contacts made by the field interviewers. The field interviewer may contact the farm operator on the telephone or in person. Completed questionnaires are returned to the office by mail (if time allows), phoned into a contact in the SSO who records the information on paper or CATI, or hand carried to the SSO if needed to meet the schedule.

Sample sizes vary greatly from SSO to SSO and quarter to quarter. The following table provides the September Agricultural Survey sample sizes for selected states as well as totals for the nation.

September 1993 sample sizes

STATE	List	NOL	Total
Colorado	1,550	150	1,730
Indiana	2,690	160	2,850
Texas	2,950	500	3,450
Wyoming	690	110	800
US	69,200	7,740	76,940

With this volume of questionnaires to be accounted for in a limited time, the SSO must carefully plan the flow of data through the office and use a tremendous effort to complete the work on schedule.

Questionnaires that are completed on paper are hand edited and then keyed in a heads down manner using key entry software, such as Key Entry 3. Heads down entry is required to meet the rigorous schedule.

A mainframe batch edit system is used to check each form for validity. Batch edits are repeated until all questionnaires are clean. All SSOs have access to an IBM Mainframe located in Orlando, Florida, on which NASS has contracted for processing time. Almost all machine editing, analysis and summary take place on this mainframe.

4. The September test

We will conduct a test of Blaise for the September Agricultural Survey in three states: Colorado, Indiana and Wyoming. The scope of the work will include conducting all CATI work in Blaise and using interactive editing (IE) on all questionnaires. All data will then be run through the mainframe edit to provide an interface with the post edit processing on the mainframe.

CASS has spent the last few months developing a prototype instrument to do both CATI and IE. The instruments should be completed by mid-August. The automated procedures to generate the instruments mentioned earlier are working well.

Much time was also spent designing the flow of data through the Blaise system to provide the SSO with the most efficient management of thesurvey. Meeting all the issues mentioned above required much planning. An overview of the data flow, as currently envisioned, follows.

4.1 Processing data flow

There will be two physical Blaise data sets. The first data set is primarily for CATI and all forms are initialized here prior to the start of the survey. Completed forms whether done in CATI or keyed from paper will be moved from the "CATI" data set to the second data set, which we will call "edit". All interactive editing will take place in "edit".

The keys of each form are the State FIPS code, the List Frame ID (list samples) or segment number (NOL samples), the tract number and subtract number. The state FIPS code is a standard two-digit identifier for each state. For NOL samples, the tract identifies the area tract from the June Agricultural Survey, while the tract is always '1' for list samples. The subtract is '1' for all forms except when an additional operation is found and needs to be added. In this situation, the first three keys are the same for the "added" operation, but the subtract is assigned a unique number, thus giving it a unique key.

The processes as numbered on the flowchart (figure 1) are:

Pre-survey

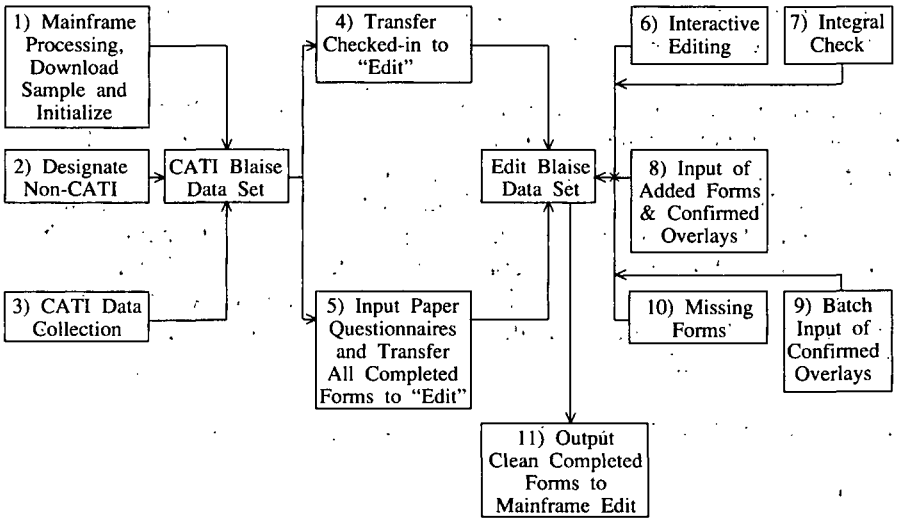
1) Mainframe processing, download and initialize in Blaise
2) Designation of CATI/Non-CATI

Survey proper

3) CATI data collection (with or w/out call scheduler)
4) Transfer Checked-In (Mail Returns and/or field completed) forms to "Edit"

5) Input paper questionnaires and transfer of all completed CATI forms to "Edit"
6) Interactive edit
7) Integral check
8) Input of added forms & confirmed overlays
9) Batch input of confirmed overlays
10) Missing forms
11) Output clean completed forms to mainframe edit

Figure 1 Blaise data flow for September Agricultural Survey



Details follow on each process:

4.2 Pre-survey

1) Mainframe processing, download and initialize in Blaise

The name and address information is maintained on the mainframe. Mainframe processing steps prepare the files for downloading. They are

then downloaded to certain directories on the LAN. The SSO then runs a couple of processes which uses Manipula and Convert to build the initial Blaise data set.

2) Designation of CATI/Non-CATI

As mentioned above, the entire sample will be set up in Blaise. There will be samples, however, that the SSO does not want to call on CATI. In some cases, due to the size of the workloads, certain samples will be sent immediately to the field and not attempted in CATI. This enables the field staff to start work while the office is doing CATI. Other forms withheld from CATI are typically those that the SSO wishes to give special handling.

All forms will be initialized as eligible for CATI, so records must be input into this step for each form to be held out of CATI. These records could be keyed or produced through some automated means by the SSO. This step uses Manipula and Convert.

4.3 Survey proper

3) CATI data collection (with or without Call Scheduler)

Virtually all CATI calls will be done using the Call Scheduler. This will be the first Blaise application where sample sizes will be sufficient to use the Call Scheduler effectively. SSOs have not traditionally had a Call Scheduler available, so we will be studying how well this works with great interest.

During the period of time when CATI is used, the SSOs will typically start phoning when the office opens around 7:00 am. The number of daytime interviewers will vary from one to ten, dependent on the size of the sample and the number of phone lines available. During the day, the regular office staff has many of the workstations on the LAN committed to tasks other than CATI. After regular office hours, a larger

staff of interviewers goes to work. This would typically be from 5:00 to 9:00 pm. The number of interviewers vary from 6 to 40.

4) Transfer checked-in (mail returns and/or field completed) forms to "Edit".

Most SSOs will have a procedure whereby they keep track of questionnaires that have been completed. The more modes of data collection that are occurring, the more important this step is. When the time comes during the survey cycle to notify the field interviewers of the sample units that are not yet accounted for, it is the check-in system that the office uses to identify them. The SSO will continue to use their existing check-in system during the September test.

The purpose of this step is to notify Blaise that a questionnaire has been received in the SSO and should not be retrieved for CATI. This step can be used to eliminate the possibility of repeating a contact with a respondent that has already responded to the survey (probably a mail return). If there is no potential for this to occur, the SSO doesnot need to run this step.

All forms will be initialized to "not in". If a check-in record is processed, a switch in the form will be set indicating the form is "checked-in". It is then moved over to "edit", thus removing any chance that the Call Scheduler could retrieve this form. This step uses Manipula, Convert and the Forms Manager.

5) Input paper questionnaires and transfer all completed forms to "Edit"

Paper questionnaires will still be keyed using software such as Key Entry 3. This data will be referred to as code/data within this document. Once a batch of code/data has been keyed and verified, it is ready for input into Blaise. This step on the data flow is actually a series of Manipula, Convert, Forms Manager and Foxpro programs. The major purposes of this step are:

A) Convert the code/data into the Blaise data set. This occurs when the code/data matches a form that had not previously been completed (the most frequent occurrence).

B) Identify potential "added forms". If the code/data does not match a form, it is identified on a listing. This may indicate that the key fields from the paper questionnaire were miskeyed or that the operation is a potential "added form". Legitimate "added forms" will be entered in step 8.

C) Identify potential overlays. Because of the multiple modes of data collection, there is the possibility that the SSO will have more than one questionnaire for the same sample unit. This could be two paper questionnaires (one mailed, one phoned) or a combination of CATI and paper. Thus if the code/data matches a form which has already been completed, the SSO is notified of a potential overlay and the code/data in question is saved for possible use in step 9.

   The editor can then evaluate which interview has the better data. Where the overlay is desired, it will be handled in steps 8 or 9. Where the old report is to be kept, the user need not do anything, since it was not overlaid.

D) The forms receiving code/data, or completed in CATI since the last time this step was completed, are then moved into a temporary Blaise data set. The creation of the temporary data set allows an integral check to be run on the data before moving the forms into "edit". After the integral check is run, the forms are transferred from the temporary data set into "edit".

6) Interactive edit

All editing will take place in "edit". We expect that one statistician will edit the entire form and correct all errors. Interactive editing will

be new to the SSOs and its impact will be very interesting to observe. The time to be saved here over the current batch edit review could be substantial.

7) Integral check

Since forms are run through the integral check prior to reaching "edit", the need for running an integral check here should be rare. This step might be run if a fix were made to an edit during the survey and it was necessary to pass all forms through the revised edit. This step might run for half an hour on the September Ag Survey. Since this step must be run when no other activity in "edit" is occurring, it will be necessary for the office to carefully schedule it.

8) Input of added forms and/or confirmed overlays

These will be identified at step 5. The office will enter "added forms" in IE. This application will allow new forms to be added in only very limited circumstances. The instrument will only allow an "added form" to be clean if it matches the first three keys (state, ID or segment and tract) of an original sample unit. "Added forms" not meeting this match criteria would have to be deleted as they could never be clean.

For potential overlays, if the state wants the original data overlaid they can do so through the IE screens. Step 9 is provided as an alternative.

9) Batch input of confirmed overlays

If after reviewing the potential overlays, there are more confirmed overlays than the SSO wants to deal with through IE, this step provides a means to use the code/data, which was saved in step 5, for the potential overlays. The SSO can use a text editor to eliminate any records that should not be overlaid. They will then run the remaining code/data for the confirmed overlays through this step, which basically is identical to 5) except that forms are overlaid.

10) Missing forms

With a large number of questionnaires coming into the office from a variety of sources in such a short period of time, it is difficult to precisely keep track of the status of each form. A procedure using Manipula and Abacus has been written to provide the SSO with a review. This step would normally only come into play very late in the survey, when the SSO is trying to wrap up the survey. The report generated here will indicate the following:

Missing Form – form which has not been completed

Good Added Form – "added form" meets minimum match requirements with original sample

Bad Added Form – "added form" does not meet minimum match requirements (must be deleted)

As long as forms are missing or bad subtracts exist, the data set is not clean by NASS standards.

11) Output clean completed forms to mainframe edit

Periodically, data in Blaise will be converted to code/data and uploaded to the mainframe for input into the traditional batch edit. Only completed forms that are clean and have never been output, or are clean but have been changed since they were last output, will be converted. These are then converted, by using Convert and a NASS developed Foxpro program, into code/data and uploaded to the mainframe. The standard edit is then run to verify that the reports from Blaise are also clean by the mainframe standards. The goal is to gradually phase out the mainframe edit.

5. Results

At the time this paper is being submitted the final steps leading up to the September test are taking place. If during August, the design of the

application changes, an update will be given when the paper is presented in London. The results from September will also be discussed at the conference, as will the impact that this will have on the future of CASIC in NASS.