Pretest 5 Report

The pretest for the proposed data capture options for Census of Agriculture and Fisheries (CAF) involved transcribing the entries of questionnaires that were ICR/OMR (intelligent character recognition/optical mark reading)-ready as well as data entry-ready. The pretest covered 200 questionnaires for CAF Form 2, and 100 each for CAF Form 3 and Form 4.

Management wanted to assess how the data will be processed electronically. Based on initial studies, the IRD proposed four different data processing options. These are described as follows:

1. Document Imaging with full interpretation/recognition

   CAF forms 1, 2, 3 and 4 will be scanned. All items /fields will undergo the interpretation stage of EHF. Only items/fields with interpretation and validation errors will be subjected to the verification stage.

2. Document Imaging with minimal interpretation/recognition

   CAF forms 1,2, 3 and 4 will be scanned. All mark fields (OMR) will undergo interpretation. All write-in fields (ICR) will not undergo interpretation, but will be subjected to 100% verification instead.

3. Document Processing thru Data Entry (imps/cs pro)

   CAF forms 1,2, 3 and 4 will be processed using the conventional data entry procedures.

4. Document Processing thru Data Entry using Images as source

   CAF forms 1, 2, 3 and 4 will be scanned. Data will be keyed-in using the images as source (Key-from-image).

General Instructions For Transcribing

1. Use black ballpen to record information.

2. Solid lines and larger boxes in the final set of questionnaires require write-in entries while smaller boxes need to be marked with "X".

3. The flow of questionnaires differed from one pretest to another. Refer to the transcription guide to determine where each item from the pretests questionnaires will be transcribed to the final set of questionnaires. The guide is by questionnaire type and by pretest.

4. There are two ways to correct mistakes in transcription and each transcriber should try doing both:

   a. If a mistake is made in writing the "X" mark in the box for the correct code, line out the wrong entry and write the "X" mark in the correct box and encircle its corresponding code.

   b. If an error is made in transcribing write-in entry, line out the wrong entry and write the correct one on the nearest space.

c. Put a sticker over the wrong marked box and mark the box corresponding to the correct answer. Put the sticker over the wrong write-in entry and write the correct entry on the sticker.

Conduct of the Pretest

a. Transcription

The data used in the test was derived from the questionnaires accomplished during the previous pretests. Since the questionnaires were not suited for document imaging technology, the first stage of the pretest was to design and produce questionnaires that could be used both for ICR/OMR document processing and the conventional data entry. The printing of questionnaires was contracted out.

Before the actual transcription, the questionnaires were subjected to a simple quality check to ensure that these will be identified during the interpretation stage. It is important to note that the inclusion of pre-printed serial numbers in the forms made checking easy. (refer to sample)

The pretest covered 215 questionnaires for CAF Form 2, 105 for CAF Form 3 and 20 for CAF Form 4. It was recommended that in transcribing the data from the old questionnaires into the final set of questionnaires, actual field conditions should be simulated. The CPOD staff who was tasked to do the transcription, adopted the following strategies:

- The assigned persons were composed of the following:

    1. right and left handed
    2. with 20/20 vision and some wearing eyeglasses
    3. young and old (not more than 45 years)
    4. with legible and illegible handwriting strokes

- The transcriber filled out the questionnaires while in the following positions

    1. Standing using clipboard, hand supporting the clipboard, with umbrella (to simulate "enumerating during rainy days")
    2. Standing using Enumerator's Manual(EN) as clipboard, hand supporting the manual
    3. Standing, writing on a window sill of a one-storey house
    4. Standing, writing on top of a sari-sari store *"escaparate"*
    5. Sitting at the top step of a wooden ladder using clipboard
    6. Sitting in a sofa with clipboard
    7. Sitting using EN Manual as clipboard on lap
    8. Squatting (to simulate "the respondent is currently washing clothes")
    9. Using questionnaires as clipboard

Also, there was a deviation from the traditional use of pencil in filling up the forms. For the pretest, black pen (PANDA ballpen) was used. This aimed to decrease the possibility of misinterpretation during data processing. Based on C2K experience, one of the reasons for high misinterpretation rate was the use of a very light handwriting strokes using pencil. After these questionnaires were scanned, the items/fields in the images produced were too light for the interpretation software to fully recognize.

b. Data Processing

The questionnaires were scanned using Kodak capture software. After scanning, the questionnaires were forwarded to IRD-FOMS for data entry (option 3). Two data entry operators were assigned to key-in 340 questionnaires. The average output of the assigned operators based on their current workload were as follows:

Operator	CPH 2000	Import
Jannette	7,747 keystrokes per hour	14,201 keystrokes per hour
Luvy	9,748 keystrokes per hour	14,901 keystrokes per hour

While the questionnaires were being keyed-in, the images produced during scanning were processed for options 1 and 2. Two verify operators from DCC-Manila were alternately assigned to do the verification process for both options. Two output files per type of form were produced after the transfer module.

To measure the time spent to process the data, statistical log files were produced. These files were then consolidated in one report. (refer to worksheet)

c. Data Comparison

The output files of the different options were compared for quality check purposes. The comparison was done on a field and character basis. The process was carried out using the following methodologies:

a. Determining Questionnaires with discrepancies

   A special program was designed to match the records in the output text files of all options using the questionnaire serial number. The data of the matched records were then compared character-by-character. The program automatically generated a report on how many characters per questionnaires did not tally.

b. Selecting Sample Questionnaires

   The questionnaires were sorted based on the number of unmatched characters. Using PPS sampling scheme, sample questionnaires were selected. Outlier questionnaires, (questionnaires having almost 100 discrepancies) were excluded in the selection process.

c. Actual Forms vs. Electronic Copy

   The actual forms of the sample questionnaires were manually compared with the text files of the different processing options. A tally sheet was used to record the total fields of questionnaires with data not matching the entries in the actual form.

d. Result of Pretest

The results of the CAF data processing pretest showed the difference between the options presented with reference to speed and quality. In terms of Speed, Option 1 – Full Interpretation proved to be the fastest processing scheme while Option 3 – Data Entry is the slowest. For quality, Option 2 produced the most reliable data with a 0.30 % and 0.12 % error rate for forms 2 and 3 respectively.