EICV2 Final Data Processing Report:

February 2007

Draft report

Geoffrey Greenwell, Oxford Policy Management

March 2007

Contents

1. Data Processing Methodology

New systems and techniques were used to capture and edit the data for the EICV-2. Many improvements were implemented to the data entry system for the EICV-2. The EICV-1 used the DOS based software called IMPS for both data entry and data editing (CENTRY and CONCOR modules respectively). In addition, EICV-1 used various short term and intermittent consultant inputs for the design and implementation of the data processing system. The first five months of the data entry process during the EICV-1 suffered greatly from a lack of quality control. This lack of cohesive support during both the design phase and initiation of the data processing system likely impacted the quality of the data despite attempts made to correct the system during mid-survey.

For the EICV-2, long-term and continuous technical support was provided by the OPM consulting firm and better trained and more committed local supervisors followed through in implementing and maintaining the system. In addition and more importantly, the EICV-2 data processing activities followed quickly behind the processing of the DHS (Demographic and Health Survey). It was clearly advantageous to simply adapt the DHS data processing system for the EICV-2. The DHS data processing system is a broadly used and dynamic system designed for use with the data processing software CSPro (Census and Survey Processing System). In fact, CSPro is designed with the DHS as its model survey. Furthermore, this system of managing the data processing activities is also being used by UNICEF to process the MICS. Applying a robust system and modifying it for use during the EICV-2 saved a great deal of time and effort in training and development. The staff was already familiar with the DHS data processing and editing system and porting the system to the EICV-2 over the long term and through the extent of the survey proved very useful. Some of the specifications that are used by the DHS, MICS and the EICV are:

a. An integrated sample design control sheet used to check in questionnaires.
b. A data entry system designed as “system control”. A system controlled application is a very tight control system where the path of data entry cannot be circumvented by the data entry clerk. The path is fully programmed and must include: skips and pre-defined keys for: missing, other or incoherent data.
c. Full double-entry for independent verification.
d. A systematic and logged passage of data files from: primary-verified-raw-edit-final data files.
e. Full reconstruction of the consolidated data file with the primary cluster file.
f. All corrections done on the lowest ASCII cluster level.

1.1 Selection and Training of Data Entry Clerks

A general announcement was posted seeking candidates to enter data for the survey. The minimum qualifications established are presented

Box 1.1 Qualifications for data entry clerks

French

Description du travail:

Les agents de saisie pour l'EICV 2 serons les agents qui reçoivent les questionnaires du terrain et seront responsable d'entrer ses donnes tels comme ils sont trouvé dans le questionnaires. Les expectatifs des agents de saisie sont les suivants:

- *Pouvoir tapez minimum 40 caractères numérique par minute*
- *Un taux des erreurs de 5% (2 sur 40 erreurs par minute maximum)*
- *Pouvoir lire des instructions en Français*
- *Montrer la capacité de suivre la logique du questionnaire et être conscient de la logique du questionnaire*
- *Maîtriser le logiciel Excel*
- *Une diplôme en secrétariat en dactylographie*
- *Avoir expérience en saisie en CSPro ou IMPS sera un atout*
- *Travailler ensemble dans une équipe*
- *Avoir réussi l'examen de saisie.*

English

Job description

The data entry clerks for the EICV2 will be the agents that receive the questionnaires from the field and will be responsible for keying the data as they are provided in the questionnaire. The following is expected from the data entry.

- Type at least 40 numeric characters per minute
- An error rate no greater than 5% (2 out of 40 key strokes)
- Follow instructions in the French language
- Follow the logic of he questionnaire
- Show proficiency in the use of Excel.
- Hold a diploma in typing
- Have previous experience in the use of CSPro or IMPS
- Work in a team environment
- Passed the data entry exam

There were 88 initial candidates that were selected from applicants. These candidates were subjected to a series of practical tests. 23 clerks were selected out of 88 applicants. These data entry recruits underwent a training session during the period of September 28-October 14 (the first four days being a training of trainers). Many of these recruits had prior experience in keying the DHS. The manuals used during the session were adapted from the DHS as well. These are available in the data archive.

The list below provides the names of the 23 selected data entry clerks. In addition to the data entry clerks, three persons were hired (as archivist) to receive, check-in, distribute and shelve the questionnaires.

Table 1.1 List of data entry personnel for EICV 2

1. Euphrosine
2. Clarise
3. Anne Marie
4. Antoine
5. Alphonsine
6. Jean Marie
7. Aimable
8. Marie Claire
9. Eugénie
10. Solina
11. Béata
12. Constance
13. Vestine
14. Phocas
15. Rose
16. Béatrice
17. Marie Josée
18. Sylvie
19. Jacqueline
20. James
21. Nicole
22. Julienne
23. Laurence

The following lists the compensation provided. This compensation was likely low as it was based on the 2000 survey with accounting for inflation. Competing surveys such as the DHS provided a great deal more. This discrepancy in compensation created some initial conflict. Pay for contracted type work should follow an institutional standard across surveys.

> The data entry personnel were paid 100.000 Rwandan Francs (gross salary) per month with a daily allowance of 1.500 Rwandan Francs.
>
> The data entry supervisors were paid a gross salary of 270.000 Rwandan Francs per month with no daily allowances.
>
> The archivist (maintenance of the questionnaires) was paid 120.000 Rwandan Francs (gross salary) with no daily allowance.

Data entry activities officially began on October 24, 2006. There were three questionnaires requiring tracking. These are respectively called Part A (demographic and person characteristics), Part B (expenditure data) and Part C (Community Questionnaire). Table 1.1 outlines the specific plan for data entry. There was not enough time provided to adequately pilot the questionnaire and the data entry system. Table 1.2 below provides a schedule of activities by cycle.

Table 1.2 Programmed schedule of data entry

Cycle	Receive Part A	Finish Part A: First Entry	Finish Part A: Double Entry	Receive Part B	Finish Part B First Entry	Finish Part B Double Entry	Receive Part C	Finish Part C Single and Double entry
1A	24/10	26/10	30/10	31/10	4/11	8/11	9/11	9/11
1B	10/11	12/11	16/11	17/11	21/11	24/11	25/11	15/11
2 A	26/11	28/11	2/12	3/12	7/12	11/12	12/12	12/12
2 B	13/12	15/12	19/12	20/12	24/12	28/12	29/12	29/12
3 A	1/1	3/1	7/1	8/1	12/1	16/1	17/1	17/1
3 B	18/1	20/1	24/1	25/1	29/1	3/2	4/2	4/2
4 A	5/2	7/2	11/2	12/2	16/2	20/2	21/2	21/2
4 B	22/2	24/2	28/2	1/3	5/3	11/3	12/3	12/3
5 A	13/3	15/3	19/3	20/3	24/3	28/3	29/3	29/3
5 B	30/3	¼	5/4	6/4	10/4	16/4	17/4	17/4
6 A	18/4	20/4	24/4	25/4	29/4	3/5	4/5	4/5
6 B	5/5	7/5	11/5	12/5	16/5	22/5	23/5	23/5
7 A	24/5	26/5	30/5	31/5	4/6	8/6	9/6	9/6
7 B	10/6	12/6	16/6	17/6	21/6	27/6	28/6	28/6
8 A	29/6	1/7	5/7	6/7	10/7	14/7	15/7	15/7
8 B	16/7	18/7	22/7	23/7	27/7	2/8	3/8	3/8
9 A	4/8	6/8	10/8	11/8	15/8	19/8	20/8	20/8
9 B	21/8	23/8	27/8	28/8	2/9	7/9	8/9	8/9
10 A	9/9	11/9	15/9	16/9	20/9	24/9	25/9	25/9
10 B	26/9	28/9	2/10	3/10	7/10	11/10	12/10	12/10
Data entry was actually completed on Friday, Oct. 20, 2006. The original target date for completion was Oct. 12. Data entry was completed within 6 days of the planned date.								

1.2 Management and Flow

The data entry was done centrally in the NISR headquarters. Activity was initiated in the old Census building in Remera on October 20. On December 16, 2006, the NISR consolidated its offices and moved the Census activities to its current location in the old MINIPLAN building. The move required the establishment of the new data entry operations in the new building and the transfer of all machinery to the building. This operation did not adversely affect the keying operations. The remainder of the survey was keyed in the MINPLAN building. It should be noted that the machines that were used for the data entry were the same machines used for the Census in 2002 . 3 new machines were purchased and provided to the supervisors.

All computers were set up in a LAN with data being copied and written to the supervisor machines and backed up daily.

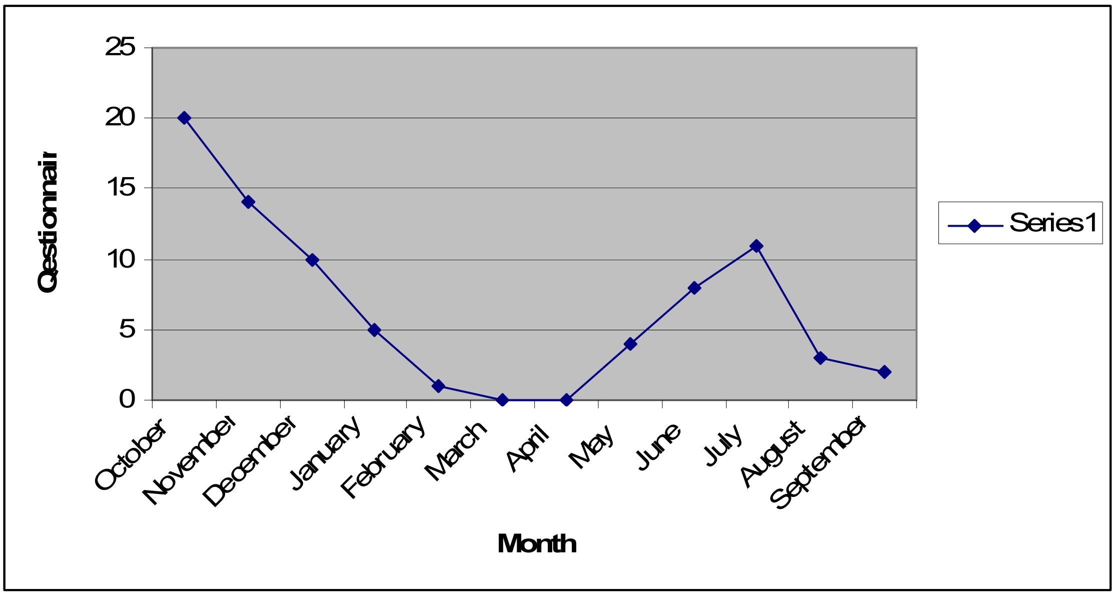
The questionnaires were received and checked into a central repository. Data was entered by the cluster (9 urban questionnaires or 12 rural questionnaires). Two archivists managed the check-in and distribution of questionnaires to the data entry supervisors. A sample of the check-in forms is provided in Annex 1.

Once the questionnaires were received and logged on a control sheet, the control sheet was entered in an automated control system by the data entry supervisors prior to being assigned to the data entry clerk. This system maintained by the supervisors assured that the sample design was strictly adhered to and that the coding and tracking of the questionnaires was properly initiated and followed. This system was built on the DHS control system and used CSPro to manage the flow and assignment of the questionnaires.

There was a 100% full independent double data entry of the questionnaires. This assured virtual certainty that inconsistencies found in the data were mostly due to errors and misreported items from the field.

For the Part A questionnaires, if they were received on time, the data entry system was designed to provide a prompt return to the field if inconsistencies were found in the questionnaire. After the completion and reconciliation of the two data files, further edits were programmed to identify further inconsistencies.

Figure 1.1 Returns of questionnaire to the field



The table

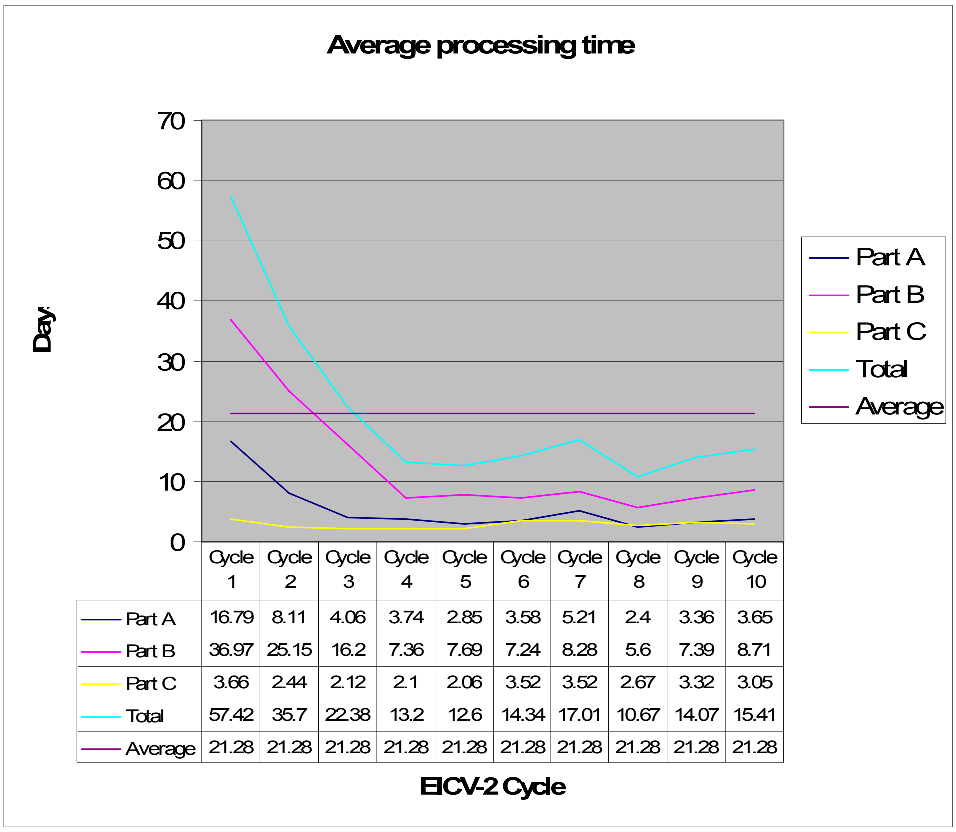
below provides the average number of days required to completely enter a questionnaire in the system This includes double data entry and reconciliation and edits.

Table 1.3 Average number of days to key the questionnaire

Part A (Household and person characteristics)	5.4 Days
Part B (Expenditure data)	13.1 Days
Part C (Community Questionnaire)	2.8 Days
Average[1]	21.3 Days

The figure below provides a graph that plots the number of days required to enter a questionnaire by the associated cycle.

Figure 1.2 Average processing time by cycle



	Cycle 1	Cycle 2	Cycle 3	Cycle 4	Cycle 5	Cycle 6	Cycle 7	Cycle 8	Cycle 9	Cycle 10
Part A	16.79	8.11	4.06	3.74	2.85	3.58	5.21	2.4	3.36	3.65
Part B	36.97	25.15	16.2	7.36	7.69	7.24	8.28	5.6	7.39	8.71
Part C	3.66	2.44	2.12	2.1	2.06	3.52	3.52	2.67	3.32	3.05
Total	57.42	35.7	22.38	13.2	12.6	14.34	17.01	10.67	14.07	15.41
Average	21.28	21.28	21.28	21.28	21.28	21.28	21.28	21.28	21.28	21.28

[1] Average days of processing includes: primary data entry, double data entry, verification and reconciliation of double entry data files.

This figure clearly illustrates a learning curve and a stabilization of the average time around the fourth cycle. Furthermore, during the first cycle, some of the questionnaires were not printed correctly and included a wrong sequencing of products. This error in coding from the questionnaire to the data entry system required a return and manual correction. The process was cumbersome and required careful attention to correct. The problem was corrected and after the first cycle, the questionnaires were consistent. There was also a slight increase in processing time in June and July. This was due to a number of errors found in the questionnaires that required correction from the field. Waiting for the return of the questionnaires increased the processing time slightly.

1.3 Primary data quality issues

Certain systematic data editing issues have been identified and should be

A. Buckets for own consumption

As part of an attempt to standardize the reporting of quantities consumed in Section 8D, own consumption (as there are a grant many non-standard units) it was proposed that a standard volume and graduated bucket be provided to the household. The bucket had 10 marks that were provided to estimate the level of consumption. However, there were delay in introducing the bucket into the field and it was not until the third cycle that it was provided.

A problem was discovered during the processing of the fifth cycle in that the price per unit reported was at times not reasonable. The field personnel were not reporting the price per standard unit which was the portion of the bucket consumed but the entire bucket. For example: if a household reported that 3/10 of a bucket were consumed, in many cases the value reported was the entire bucket and not 1/10 of the bucket. This meant that in many cases own-consumption was an order of magnitude more than expected. A routine to correct for this problem was developed. The file that is provided for public dissemination is the corrected file.

B. Problems in reporting ag activities

C. Reporting replacement value instead of current (or salvage value) of assets.
A standard method of computing depreciation values is to compute the difference between purchase and residual values (adjusted for inflation). The question is formulated such that the household responds to the purchase value and the current worth of the durable good. The instructions provided to the enumerators (erroneously) was to report the replacement value. Computing a depreciation schedule would only have produced an estimation of inflation for that specific item. For this reason, there was no computation of the depreciation rate. Instead, the same rates were applied to compute the use value for durable goods based on what was used for EICV-1 (the first survey).

1.4 Recodes and comparability

In order to assure some comparability between EICV 1 and EICV 2, variable names and category types require recoding. In order to produce the comparative study undertaken by OPM, it was necessary to develop a system of recodes. A combined data set was developed where all items were recoded to EICV-2 names and codes. However, this is not a desirable solution and the data set is not recommended for public dissemination. The recode syntax is provided, however. Furthermore product codes are generally sequential and have not been recoded to a standard (such as COICOP or NISR internal definitions). At the time of archiving the data, there was no internal policy on internal coding of products. For the purposes of the OPM study, product codes were expressed using the EICV-1 coding scheme.

1.5 Summary specifications

Software used:	CSPro 2.6
Network:	20 machines divided into two groups of 10.
Data entry machines:	Pentium 4 150 GHz with 128 MB RAM
Data entry components:	Questionnaires: PartA.ent, PartB.ent, PartC.ent Management : Centry.ent , Superv.ent

1.6 Recommendations & Observations

A more comprehensive look into the own-consumption problem regarding the buckets with a careful review of the correction scheme

Product codes should be recoded to either an internal standard or COICOP. This should be done for EICV-1 and EICV-2. A general recode file should be kept as a standard file.

EICV-1 should be completely re-documented. The current version on CD-ROM is incomplete. The CDROM produced for the archive contains a version of the dataset kept by the world bank

Other surveys should be documented.

Additional edits on agriculture

Additional edits and treatment of income.

Proper storage of questionnaires.

Develop and maintain a standard weights and measure data base.

Adopt STATA as the institutional software platform for analysis (drop SPSS).

Comprehensive archiving of all surveys.

1.7 Appendices

1.8 Control sheet for questionnaire A & B

A	**<===Partie A,B ou C**	**Province**	**District**	**Secteur**	**ZD**
	<===Numéro de grappe				

Dates:		**Ménages**	**Total Personnes**	**Total Membres**	**Membres 6 ans et plus**	**Membres Femmes 12-49**	**Membres 5 ans et moins**	**Membres 15 ans et plus**	**Commentaire (Remplacement)**
Date de Réception									
Date de Assignation:									
Agent de Saisie:									
Ordinateur									
Date de double saisie:									
Agent de Double Saisie:									
Ordinateur									
Date de correction 1									
Date de correction 2									
Date de correction 3									
Date Fin									

B	**<===Partie A,B ou C**	**Province**	**District**	**Secteur**	**ZD**
	<===Numéro de grappe				

Dates:		**Ménages**	**Enqueteurs**	**C. No de champs**	**D. No. de cultures a grande echelle**	**E. No. de culture a petite echelle**	**H. No. de produites tranformes**	**9B. Depenses Alimentaire 1=oui 2=non**	**9D Autocon 1=oui 2=non**	**10A. Tranfert effectues**	**10B. Transferts Recu**	**11A. Credits**	**11C. Epargnes**
Date de Réception													
Date de Assignation:													
Agent de Saisie:													
Ordinateur													
Date de double saisie:													
Agent de Double Saisie:													
Ordinateur													
Date de correction 1													
Date de correction 2													
Date de correction 3													
Date Fin													