SPC/ RPATW /Working Paper 2.2
7th September 2007

ORIGINAL: ENGLISH

SECRETARIAT OF THE PACIFIC COMMUNITY

REGIONAL POVERTY ANALYSIS TECHNICAL WORKSHOP
(Noumea, New Caledonia, 22 September 2007)

"Translating theory into practise in the Pacific context"

EDITING AND AUDITING HIES DATA

THE ISSUE

1. The editing and auditing of HIES data is a crucial stage of data processing that, too often, is not given the attention that it rightly deserves by survey planners and project managers. The reality of processing large amounts of survey data in a short period of time is that, unless properly and carefully managed, the quality of data will be less than acceptable at international standards and may not be fit for the intended purpose of the data collection. Technical assistance to countries for editing HIES data are usually between one and two weeks, though there have been occasions, when funding permitted, where the period of assistance has extended far beyond this time. The issue for subject matter experts to decide is the standard to which the data needs to be edited, and the resources necessary, to meet the required standard. Also of importance is what additional modifications/edits may be required to be made to the data in order for it to be fit specifically for poverty analysis.

MAIN APPROACHES

2. For HIESs where the Statistics and Demography Programme (SDP) provides technical assistance, there are five types of editing performed during the data processing phase of the survey.

- Questionnaire completion
- Data verification
- Consistency checks
- Outlier analysis
- Data audit

3. **Questionnaire Completion**: Checking that the household questionnaires and diaries are completed correctly is an important stage of input editing performed by a team of data editors. Where there are incomplete questionnaires/diaries or missing responses, decisions need to be made by the editors as to whether to accept the questionnaire for further data processing or refer the questionnaire back to the field supervisors for correction. Given the limited resources and tight timeframe, it is often difficult to refer questionnaires back to the field operations staff. In some cases, decisions are made to reject a questionnaire and eliminate the household from the survey, and in other cases, decisions are made to manually impute values based on other information in the questionnaire or from other similar household questionnaires. For example, if only one week of the diary is completed, expenditure for the second week is often imputed from items purchased or produced during the first week. This latter practice is somewhat error prone as it is often not known whether the household didn't get around to filling in the second diary or the household simply didn't purchase or produce anything during that week. Even in those cases where it is known that household didn't complete the second diary week, the expenditure contained in the first diary week may not be an accurate reflection of the second week's expenditure.

4. **Data Verification:** It is a generally accepted practice in household surveys to conduct a verification process during the data entry operations. In some cases, all data is entered twice into the data entry system to check for incorrect keying of responses. However, 100 percent verification of survey data is both time-consuming and resource intensive. Recent experience suggest that, depending on the level of error found in the data entry operations, verification can be performed on between 5-10% of the survey data. This level of verification provides estimates of data keying error for each data entry operator that enables additional training to be provided where the level of error is unacceptably high (note: this requires prompt verification output and analysis to enable the additional training to be effectively implemented). When combined with interactive editing procedures, data verification provides important information for the assessment of non-sampling error. Unfortunately this activity has been heavily neglected in recent times throughout the region, due to limited resources and strict deadlines.

5. **Consistency Checks:** To ensure that the household survey data is valid, it is necessary to check that responses are internally consistent within and between questionnaires. These checks can be performed during and/or after data entry operations through the use of interactive editing and batch editing procedures. Data entry supervisors and operators are trained, not only in the procedures for keying data, but also in the procedures for editing incorrect or missing data. These procedures include instructions for correcting or imputing values for inconsistent responses. The use of interactive editing procedures during data entry enhances the quality of the data by ensuring that only valid data is entered into the system. Combining data entry and editing functions also enables a more efficient process whereby the manual handling of questionnaires is kept to a minimum.

6. **Outlier Analysis:** For the HIES survey the analysis of extreme values is an important process for ensuring that the micro data at household level is accurate and complete. Once the data entry has been completed, the values of income and expenditure items in both the household questionnaire and diaries are analysed by comparing unit prices for each item with the prices paid by other households. For diary items this comparison requires the accurate recording of weights and number of items as well as the units of measurement. Where the values for items are missing, an average price is imputed based on the weight and number of items and the location of the household. Very large or small values are checked with the questionnaire/diary and corrected where necessary. Some experts suggest that adjustments should be made to the frequency weights of diary items where the values are atypical of household spending (e.g. custom events).

7. **Data Audit:** After the sample weights have been applied, the aggregated income and expenditure of major items is compared to external data sources to ensure the reasonableness of the HIES estimates. The income estimates are compared to government payroll figures, superannuation payments and with income tax data where available. The expenditure estimates are compared to the cost of imported items, domestic revenue of utility companies (e.g. water, electricity, gas and telephones), and revenue received for government services, such as education and health. These comparisons provide useful information for the general assessment of data quality of the survey. In addition, data users may use the information to make adjustments to HIES estimates depending on the specific use of the data (e.g. poverty analysis).

8. The above approaches are the standard approaches adopted by the SDP for HIES editing, for purposes such as the general HIES analysis and the re-basing of the CPI. They do not include any additional data editing or manipulation which may be required prior to undertaking poverty analysis. As it currently stands, staff from the SDP are not aware of what considerations may be required during the data editing/cleaning phase of a HIES in order to meet poverty analysis needs. These considerations need to be established and documented for future HIES editing where poverty analysis is to be undertaken.

RECOMMENDED APPROACH

9. It is recommended that the general approach for editing and auditing of HIES data include procedures for ensuring the completion of questionnaires, data verification, consistency checking, outlier analysis and data audit with external sources.

10. It is recommended that the requirements for editing of the dataset beyond the general approach discussed above for the specific purpose of poverty analysis be articulated. For such activities an agreed approach/set of procedures needs to be established and documented. If the dataset is adjusted in some way specifically for poverty analysis which may not be suitable for general HIES analysis, the adjusted dataset should be kept separate from the original.