Working Paper 15

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Budapest, Hungary, 14-16 September 2015)

Topic (ii): Managing and supporting changes related to editing and imputation

Getting commitment to a new editing strategy

Prepared by Felibel Zabala, Statistics New Zealand, New Zealand

I. Introduction

1. Statistics New Zealand (Statistics NZ) has an excellent reputation for the overall quality and integrity of its official statistics. In mid-2004, the department launched a strategy to streamline and standardise its business processes, methods, and systems throughout the statistical production cycle. Since then, the development of standards and progressing standardisation has become an important initiative of the department as a means to continue to produce a wide range of statistics efficiently, effectively and to a high level of quality and relevance.

2. 2011 saw Statistics NZ embark on a 10-year extensive work programme, Statistics 2020 Te Kāpehu Whetū, to change the way official statistics are produced in New Zealand. A key component of this is on-going development of processing platforms to perform standardised end-to-end processing. This provides a system that allows users to integrate data editing and imputation methods, processes, and standards, increasing the efficiency and quality of our outputs. The Household Processing Platform (HHP) was developed to process social survey data.

3. The Household Economic Survey (HES) Income is currently migrating to the HHP. Migration of HES Income into the HHP provided opportunities to implement methodological improvements, including changes to data editing and imputation processes. The current HHP has no functionality to edit data. HES is the first survey to migrate to the HHP that requires editing.

4. To ensure migration goals are realised, the support and cooperation of the HES Income editing staff was seen as crucial. This included engaging with them as early as possible in the migration process. Top-level management supported a redesigned editing strategy.

5. This paper describes the processes we undertook, which included conducting workshops with the HES Income editing staff to obtain agreement on the required editing process. We also committed to provide support and training in understanding the new methods. The paper concludes with the lessons learnt and how we plan to apply these to future system migration.

II. Household Economic Survey

6. HES Income is one of three survey vehicles in Statistics NZ's Integrated Household Survey programme. Each vehicle is centred on a major theme – HES Income focuses on income and expenditure. HES Income is an annual survey. Each year the HES survey includes the HES Income as its primary content and has regular modules such as the HES Expenditure. HES Expenditure is

conducted every three years. Another regular module is the material well-being, which is collected every year.

7. One of the main differences between HES Income and HES Expenditure is the extent to which expenditure data is collected. HES Expenditure collects expenditure information, including purchases of food, clothing, and household items. An expenditure diary is used to collect some of this detailed information in addition to the questionnaire. In HES Income, the only expenditure data collected relates to specific housing costs, including rent, mortgages, property rates, and building-related insurance. The diary is not used (Statistics NZ, 2014). Both HES Income and HES Expenditure have the Material well-being module.

8. Except for the diary, which collects data using a self-completed questionnaire, data is collected using computer-assisted personal interviews. Data collection is spread out throughout a year.

9. HES Income and its supplements provide a comprehensive range of statistics relating to personal and household income as well as household expenditure. Information from HES Income is used to support policy development and decision making by government agencies. Information from the material well-being module is used to publish selected results for life satisfaction levels and adequacy of income to meet daily needs. Unit record datasets are also key outputs required by internal and external key clients.

10. The main objectives of HES Income and its supplements are to:
- contribute to the reweighting of the consumers price index (CPI);
- supply expenditure statistics for use in estimating gross domestic product, and
- provide an indication of the overall living standards of New Zealanders.

Migration of HES into the HHP started in 2012. This provided opportunities to implement improvements to HES. These included streamlining and simplifying the HES questionnaires in recognition of the growing demand placed on survey respondents. Extensive methodological improvements, including changes to data editing and imputation processes, started in 2014. Improvements to the imputation process of personal income are discussed in another paper titled, "Let the data speak: Machine learning methods for data editing and imputation." This paper will be presented by Amanda Hughes.

II. Redesign of HES editing

11. The aim of the redesign of the HES editing process was to deliver increased efficiencies and improve process quality while maintaining output quality. Current micro-editing tasks are carried out by statistical processors and statistical analysts from the HES team as well as another subject matter area, the Prices team. The Prices team are primarily responsible for producing price indexes including CPI. Top management were specifically interested in an editing system that reduces manual editing, ie, increase automated editing as much of this as possible. Manual resolution of edit failures, including outliers, should be done by the processing team, rather than statistical analysts. The expectation was statsitical analysts should receive a clean dataset to start their data validation. Validation carried out by statistical analysts should not be focused on micro-level details, except under exceptional circumstances where this is proven to be the cause of higher level data issues.

12. The first part of the redesign of the HES editing process involved reviewing the current editing process. We reviewed documented edit rules and business processes currently in place to understand and analyse the amount of manual editing being carried out as well as how edit failures were being resolved. We also looked at edited and unedited data from previous HES years. This was a challenge because data stored, especially those during a HES expenditure year, stored edited data, rather than data originally collected by field interviewers.

13. We sat with a statistical processor supervisor to observe and acquire information on the following:

- the amount and types of edit failure detected
- the methods available and accessed to resolve edit failures.

14. We also interviewed statistical analysts from the HES and Prices teams to determine their involvement in the above editing tasks. The aim of this exercise was to determine if there were systematic ways used in resolving the manual edits. This will inform the amount of automated editing we can introduce. We also wanted to know if we can assign micro editing tasks solely to the statistical processors, including resolution of outliers.

15. Observations and interviews revealed that edit failures resolved by statistical processors were detected by range edits that cannot be adjusted and have not been revised since 2006. These were resolved by looking at related information available from a record. There were instances when statistical analysts had to be consulted. Meanwhile, statistical analysts printed out reports on the top and bottom five values of either an income or expenditure category and resolved if an outlier is errorneous by comparing these with previous periods of HES. Other external sources were consulted, where available, to resolve potential errors. Erroneous values that were manually treated were not flagged. Furthermore, no documentation existed to explain how an erroneous value was resolved.

16. Using the above findings, we drafated recommendations for an editing strategy for HES. Initial recommendations included:

- Outlier detection at the diary entry phase should focus on detecting very extreme values since undetected outliers should be picked up at a succeeding editing phase. Range edits to detect outliers should be reviewed and updated regularly. The processing team should have the ability to update these edits where needed, in consulation with the HES team.
- Use income data available from administrative data sources to inform regular updates to income range edits.
- Detected outliers for resolution should have related information, eg personal or household demogrphic information, made readily available to the processing team to enable them to facilitate sound decisions when treating the suspicious values.
- Use selective editing to prioritise outliered responses. Manually edit outliers that have significant impact on published statistics. All other errorneous outliers should be automatically treated using imputation.
- Use Banff to detect outliers rather than scanning through reports on top and bottom five outliers of an expenditure or income category. This should be done once nine-month data has been collected to ensure sufficient data is available for analysis. This coincides with the data validation carried out by the HES team on preliminary estimates produced from data available from the first nine months of collection.

17. The above recommendations received mixed reactions. There was hesitation in the use of Banff to detect outliers, justification for which was the time needed to detect and resolve these outliers did not consume much of a statistical analyst resource. Although the HES team appreciate the importance of automated editing, reluctance was around its immediate implementation. The reaosns behind this include:

- The stored diary data contains only data that have already been edited. A good imputation method relies on having true data available to simulate “missingness” or “erroneous values” and test the imputation method on the simulated missing and erroneous data.
- It is difficult to pull out information from the current databases and put these together to extract relevant information. There is now potential for applications of the R-software to

facilitate this. This will pave the way to perform analysis on past data to identify appropriate methods to treat erroneous values, thereby introducing and increasing automation.

- Absence of audit trails precludes information on how to automate treatment of detected errors. Investigations carried out using simulation of missing data is not possible since we don't have accurate data on which variables are legitimately missing because of routing.
- Manual treatment of erroneous values needs to be recorded to provide information on how similar errors may be treated in the future.
- Although a vast amount of manual edit checks are currently being done that result in making no changes, more information on whether the amount of edit rules available to detect errors and how edit failures are being resolved is required before moving to automatic treatment.

18. To resolve the differences in the recommendations and get buy-in to a new editing strategy, a two-day workshop was held between Statistical Methods and the HES team. The ultimate outcome of the workshops was to resolve the kind of editing the subject matter area needed. The first workshop was a Kaizen problem-solving workshop. Kaizen is the practice of implementing continuous improvement in every component of a production process (Imai, 1986). This was originally introduced by Masaaki in his book *Kaizen: The Key to Japan's Competitive Success in 1986*. Kaizen problem solving uses root-cause analysis to solve a problem to ensure the problem never occurs again (Kaizenworld).

19. The one-day Kaizen problem-solving workshop discussed in detail the editing processes that the HES team typically carried out. The worshop started by identifying HES key customers who were supplied unit record data. The HES team aim to provide data that is error-free since these customers query the organisation on detected errors. The HES team feel providing data with precision less than 100% affects the reputation of the organisation.

20. The Kaizen workshop revealed that manual editing led to only 5% changes to suspicious data detected by range edit rules. This led to discussions on identifying the possible causes of this undesirable outcome. Details on the number of suspicious values detected by survey year including the action taken on them are provided in Table 1. The action taken, "Edited", corresponds to a change in the suspicious value. Survey years in greyed boxes indicate HES expenditure years. There was no indication on how much of the potential outliers checked led to manual changes.

Table 1. Number of suspicious values detected by survey year and action taken.

Action taken	Survey year								Average
	06/07	07/08	08/09	09/10	10/11	11/12	12/13	13/14	
	Number of suspicious values								
Passed	1,628	239	179	879	160	177	870	209	543
Edited	68	16	10	34	27	9	45	27	30
Total	1,696	255	189	913	187	186	915	236	572
Error hit rate	4.01	6.27	5.29	3.72	14.44	4.84	4.92	11.44	5.16

21. The Kaizen workshop ended with identifying root causes to the following:

- Doing what we've always done
- Why are we constrained by the system design?

It was agreed we would hold another workshop to work through the roles and responbilities between the statistical processors and the HES team with regard to the editing process. This session would help inform the HES team on what and when editing is required. The desired outcome would be valuable information to decide which team should be responsible for micro editing.

22. The follow-up workshop focused on three key stages of the editing process where eventual changes to the data are carried out: daily data processing, data validation or resolution of potential outliers, and data validation immediately before the release of data. The workshop started by identfying the following with regard to each of the key stages of the editing process:

- the current tasks completed under each stage
- the outcomes required for each stage
- the roles and responsibilities required for each stage
- the system requirements for each stage.

23. The recommendations from the workshop centred around three areas:

a) Formalising the relationship between the HES and processing teams. This entails formalising what is expected from the processing team by the HES team and vice-versa as well as formalising the process for handing data over from the processing team to the HES team for analysis.

b) Building the relationship between the HES and Prices teams. This includes looking for opportunities for the two teams to work together during the data validation phases to reduce the number of queries raised by the Prices team, minimise duplication of editing tasks between the two teams, and consider the possibility of co-locating the two teams.

c) Developing a "How-to guide for data validation." This should serve as a useful resource for new people working on HES editing and define clearly the roles and responsibilities of people carrying out HES editing tasks.

24. The outlier detection procedure in Banff was demonstrated to the HES team to get buy-in. A standard procedure to detect outliers was needed to ensure this process is consistent, transparent, reproducible, and removes subjectivity.

25. Using the recommendations from the workshops as well as some of the initial recommendations, a HES editing strategy was drafted.

III. The HES editing strategy

26. The HES editing strategy provides the plan for the recommended editing system for HES to reduce the level of manual editing without compromising the quality of the datasets and outputs produced from the survey. The editing strategy provides the plan to achieve this. The strategy applies to HES Income as well as its supplementary modules. Moreover, as stipulated in Statistics NZ's *Methodological Standard for Editing and Imputation* (Statistics NZ, 2010), the proposed HES editing system should:

a) provide users with fit-for-purpose, plausible data and outputs by the most effective and efficient means
b) ensure all users are well informed about the quality of the data and statistical outputs
c) continuously improve our end-to-end business processes and overall data quality.

27. These objectives are supported by the following key principles:

a) Statistics NZ should maintain, wherever possible, the original data provided by the supplier. Enhancements to the imputation process ensure this principle is supported by HES.
b) Ensure that anticipated or likely errors are considered during development or re-development of statistical outputs and, where the potential impact of these is significant, put processes in place to eliminate or reduce the impact (ie ensure quality by focusing on error prevention at the design of the HES questionnaires). The HES questionnaires had been redesigned so that anticipated or likely errors were considered and BLAISE edits developed to reduce, if not eliminate, these errors.

c) Choose editing and imputation methods that support the two main uses of the data (aggregates and microdata analysis). For example, choose imputation methods that preserve the relationships across related variables or records.
d) Attempt as much editing as is practicable at the point of contact with respondents (eg via CAI tools) to realise the maximum benefit from the initial contact.
e) Automate the editing and imputation process, where possible, ensuring the best possible use of editing resources, for example, minimising manual intervention. The existing editing approach for HES data was mainly manual and targeted respondents with extreme values.
f) Keep an editing and imputation audit trail of the following things:
    - unedited and edited data, so that sources, types, and distributions of errors can be monitored
    - unimputed and imputed data, so that the degree, methods, and sources of imputation can be monitored
    - production, monitoring, and analysis of key editing and imputation quality indicators in order to evaluate and understand the editing process, including its cost effectiveness and efficiency.
g.) The software application should support continuous improvement, and be flexible and configurable by users to enable future developments.
h.) Wherever possible, new developments should use an editing and imputation tool in the Statistics NZ set of standard tools.

28. The following two flow diagrams summarise the editing processes for HES. Figure 1 summarises the editing process carried out on variables from the income and expenditure questionnaires while Figure 2 presents the editing process carried out on variables from the expenditure diary.

Figure 1. Editing process carried out on variables from the income and expenditure questionnaires.

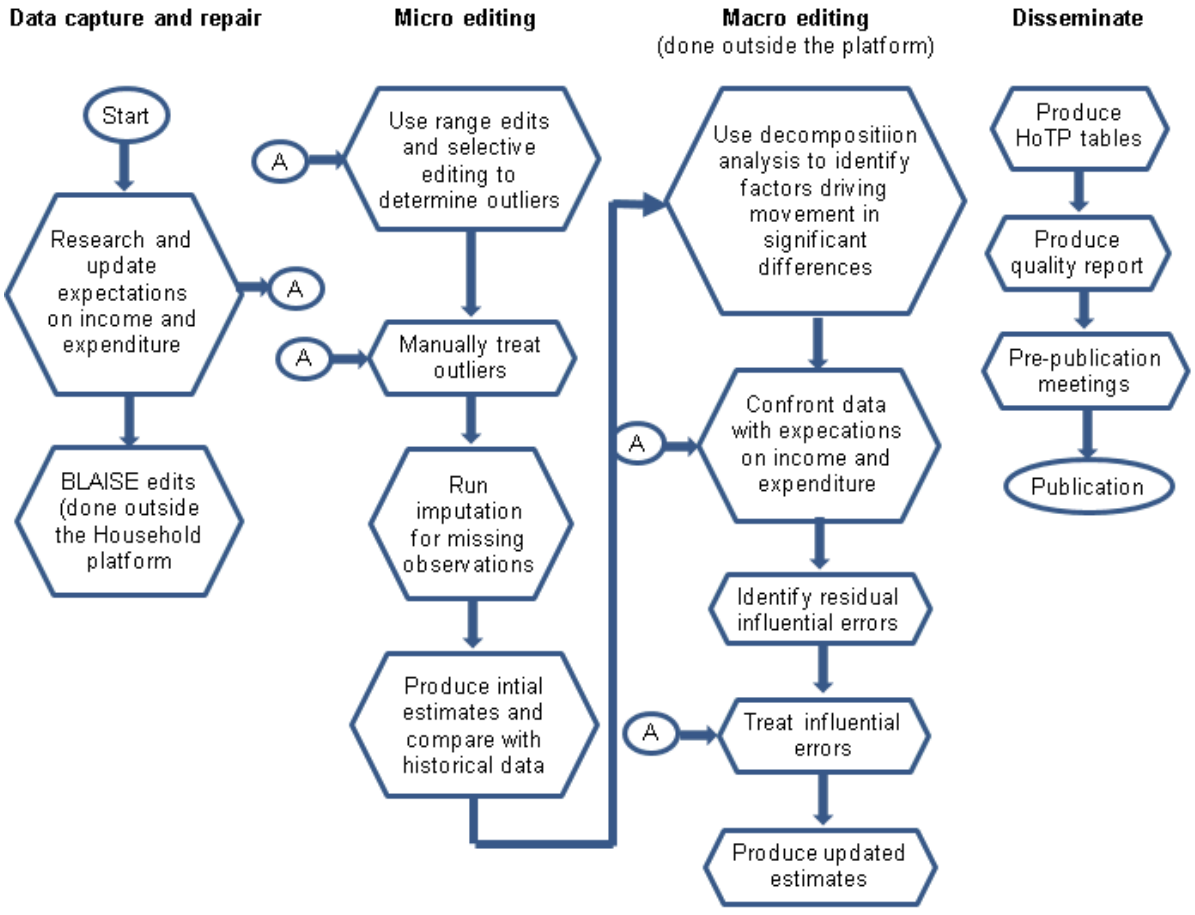
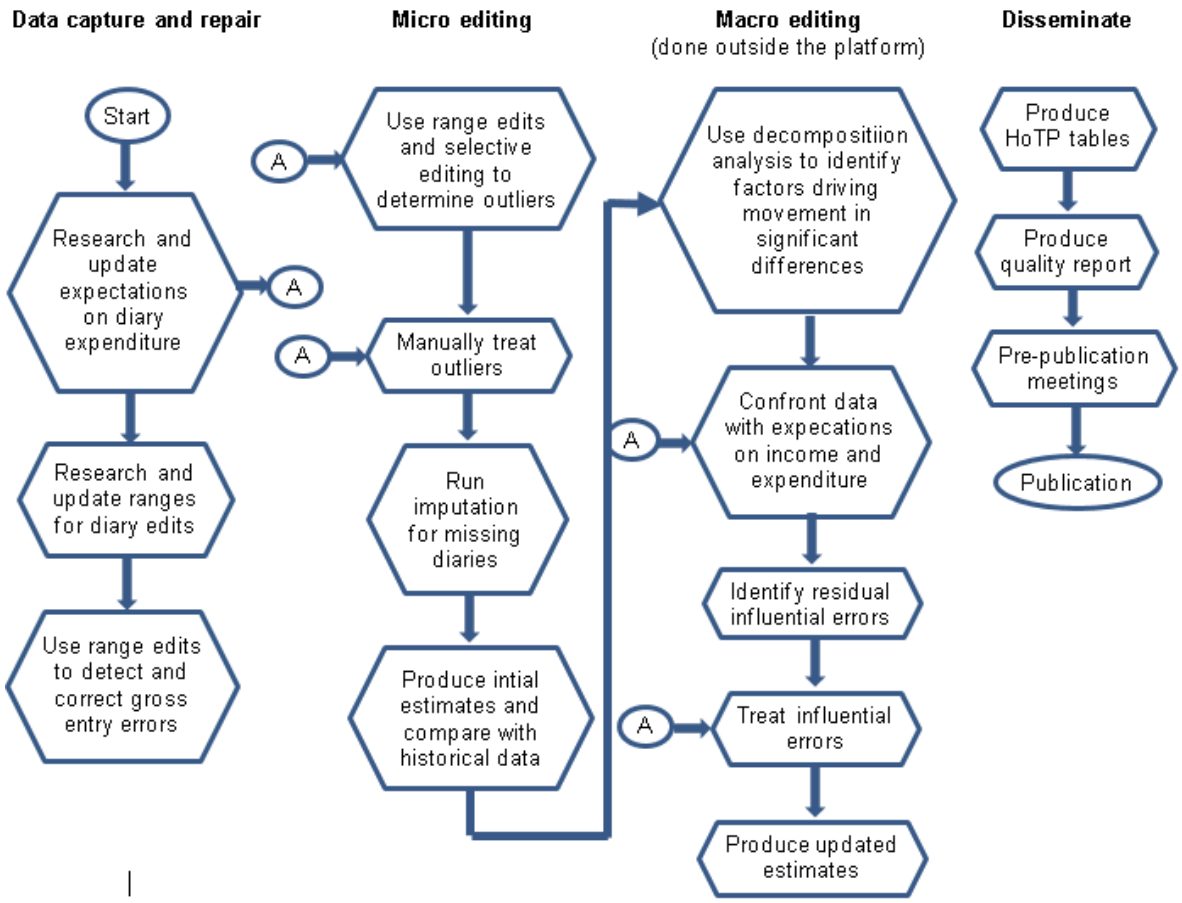


Figure 2. Editing process carried out on variables from the expenditure diary.



29. Canceis and Banff are Statistics NZ's standard tools for editing and imputation. The editing system for HES aims to make the best use of these tools. Canceis is designed to edit and impute both numeric and categorical data while Banff is designed to edit and impute continuous numeric data so is most useful for processing HES data. Both tools preserve as much of the original data as possible and provide clear editing and imputation audit trails. Moreover, Banff provides a good range of editing and imputation methods. An efficient use of both tools, together with some additional SAS functionality, can provide an end-to-end editing and imputation process that is automatic, repeatable, and in which the quality is controlled via parameter settings. Canceis is available in the HHP. Banff is the editing and imputation tool used in Statistics NZ's micro-economic platform, or MEP. Banff will be made available in the HHP to edit and impute HES data.

30. The Nearest-Neighbour donor imputation method available in Canceis will mainly be used to impute missing variables from the income questionnaire, some variables from the expenditure questionnaire, and a missing diary for a person belonging to an eligible responding household. This method is recommended since most of the matching variables used to determine the nearest neighbour donors for imputation are categorical. Banff will be used for editing as well as for imputing some expenditure variables.

31. To ensure editing feeds into continuous improvement of our end-to-end business processes and data quality, the following should be developed:

a) editing and imputation audit trails including flags for edited and imputed vales
b) reports that will enable production of key quality indicators for editing and imputation
c) storage of versions of edited and imputed data.

32. Agreement from the HES team was obtained before the strategy was presented to management for sign-off. Statistical Methods commit to provide training to staff undertaking editing and imputation, where needed. Agreement has also been achieved that the strategy will be evaluated at the conclusion of HES 15/16. The evaluation should determine the extent the editing strategy has enabled the production of quality outputs as well as the effectiveness of the editing and imputation methods. Conclusions drawn from this evaluation should feed into improvement of future iterations of HES Income.

IV. Next steps

33. System requirements for editing, based on the editing strategy, have now been defined and are currently being designed. Development of these editing requirements for the HHP ensured these satisfy requirements of other social surveys migrating to HHP that would require editing.

34. We are currently exploring statistical techniques to facilitate identification of variables that will assist statistical processors in resolving suspicious errors. We envision that results and findings from these investigations will feed into future development of an automated editing system for HES.

35. Future work includes determining the most efficient and effective macro-editing method(s) to enable identification of residual influential errors in the data. Since HES produces a vast number of key variables, graphical editing is preferred to perform macro editing. It is also imperative that interactive editing be introduced to facilitate drilling down of high level aggregates to detect residual influential errors in the data.

36. Capture of audit information will be designed for editing HES15/16 data. All changes to the data, whether system or manually driven, will need to be fully documented and will include the reason for a change. Relevant flags will be set for each action taken. Information derived from the audits will be used to evaluate the amount of automatic editing to undertake in the next iteration of HES. The ultimate aim is to minimise time micro analysing the data and focusing more on trend analysis and telling stories from our data.

V. Challenges and lessons learnt

37. Introducing changes to the editing culture where there is strong emphasis on perfection has not been easy. Future introduction of editing and imputation audit trails as well as the production of key quality indicators for editing and imputation will be useful to understand and manage survey costs.

38. Another challenge has been for processing staff to understand the connection between the data processing they carry out and the goals of the organisation which are automation, standardisation, and continuous improvement. Editing staff need to recognise the benefits of these goals such as consistency, transparency, and reduction, if not complete removal of subjectivity, in carrying out editing tasks.

39. When designing an editing strategy, it is important to understand the root causes for the need to do editing. This will provide useful information in determining the editing requirements and prioirties of a data collection as well as how future editing will look for the data collection. This facilitates agreement and support to a new strategy.

V. References

Imai, M (1986). *Kaizen: The Key To Japan's Competitive Success,* McGraw-Hill Education.

Kaizenworld, *Problem Solving*, Available from http://www.kaizenworld.com/kaizen/problem-solving.html.

Statistics New Zealand (2014). Household Economic Survey (Income): Year ended June 2014. Available fromwww.stats.govt.nz. .

Statistics New Zealand (2010). *Methodological standard for editing and imputation*. Unpublished report, Statistics New Zealand.

VI. Acknowledgements

The author wishes to thank the following members of the HES and processing teams for their valuable contribution to the development of the editing strategy: Hazel Kale, Walter Moes, Jacinta Coe, Caroline Brooking, Tului Sola, and Sharon Snelgrove. The time and thoughts they had put into discussions had been instrumental in the development of the editing strategy.