

Work Session on Statistical Data Editing
(Prague, Czech Republic, 14 to 17 October)

Item 4 of the provisional agenda

**A SMALL STUDY ON USING EDITING PROCESS DATA FOR
EVALUATION OF THE EUROPEAN STRUCTURE OF EARNINGS SURVEY**

Submitted by Statistics Sweden ¹

¹ Prepared by Per Engstrom.

I. INTRODUCTION

1. The European Structure of Earnings Survey is carried out every fourth year in the EC member countries. Sweden did the survey for the first time in 1995. The observation units are employees employed in November 1995. The information on the employees is collected from the employer. The total number of completed questionnaires, over-coverage excluded, was 16,166. The questionnaire holds 32 questions.

2. When planning the survey it was decided that data on the editing process should be collected in order to improve the next survey with respect to questionnaire design and performance of the editing process. The indicators used in this paper are proposed in Engstrom (96). One goal of the study is to see whether these indicators are adequate.

3. The following issues are discussed: the usefulness of data on the editing process, what indicators and other data are useful, and how should the indicators and other information be presented.

II. THE EDITING PROCESS

4. In this survey scanning and optical character recognition (OCR) were used. Typically the use of OCR can produce some interpretation errors, therefore some checks are carried out at this stage. Of course, at the same time respondent errors are detected, but these are not corrected during this process.

5. This study gives only the results from micro-editing. Later, it is planned to carry out a study on the scanning process. Two major differences between this and other surveys are that firstly, the cause of error is collected (for instance misunderstanding, typing errors etc.) and secondly, it is noted whether re-contact to the respondent has taken place. Unfortunately, the error coding was not detailed or precise enough. For instance, it is not evident which variables are erroneous.

6. The experience from the coding of error causes indicates that it would be better to draw a sample of the flagged, and followed up, units and carry out a thorough, high quality coding exercise on this sample.

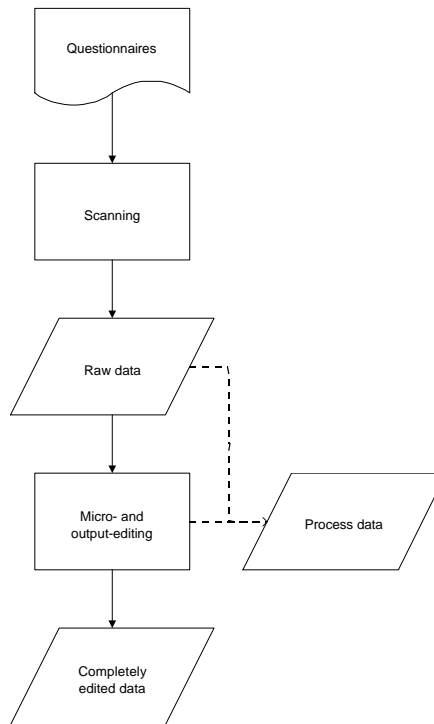


Figure 1: The process flow

III. SOME RESULTS

7. When interpreting the results, one should remember that the survey was carried out for the first time and that the questionnaire was not properly tested.

8. Overview data

The total number of flagged units were 10975, which means that 68 % of all units were flagged. Figure 2 presents a chart on distribution of flagged units by number of flags.

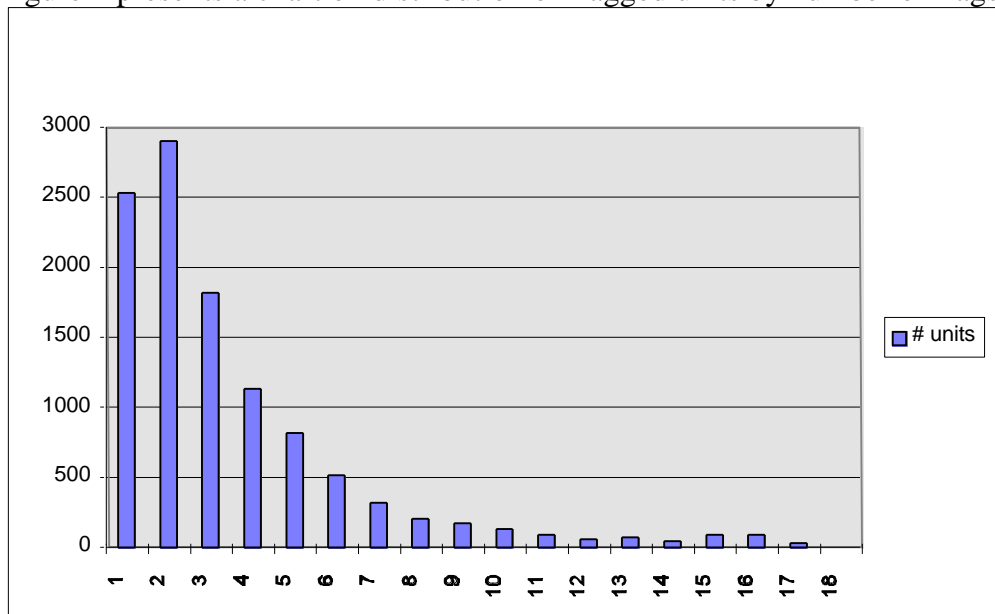


Figure 2: Distribution of flagged units by number of flags.

9. A total number of 4 598 (28 %) units were followed up for editing. The number of recontacted respondents were 2 348 (the employer can respond to more than one employee). Evidently, a lot of resources spent on editing could be saved by reducing errors made by the respondents and also by making the edits more efficient.

10. To reduce respondent errors, improvements of questionnaire design and instructions to the respondent are the most crucial tasks. One way to identify which questions to improve is to produce data on flags and cause of error, by edit and variable. These data are presented in paragraph 9.

11. The proportion of changes to the flagged records is 68 %, which indicates that the edits are efficient and do not seem to be the main problem. The proportion of changes to the followed-up records is 94 %, which indicates highly efficient follow-ups. This figure seems almost too high. One explanation could be that many telephone contacts have not been registered.

12. Data at variable/edit level

Figure 3 shows the number of flags by edit for the 10 edits which have flagged most frequently. The "V" in the edit code stands for fatal edit. The "R" in the edit code stands for query edit. Note that 9 out of 10 of the edits which have flagged most often are query edits. One way to see how efficient the edits are is to inspect a chart showing how many flags there are before and after editing.

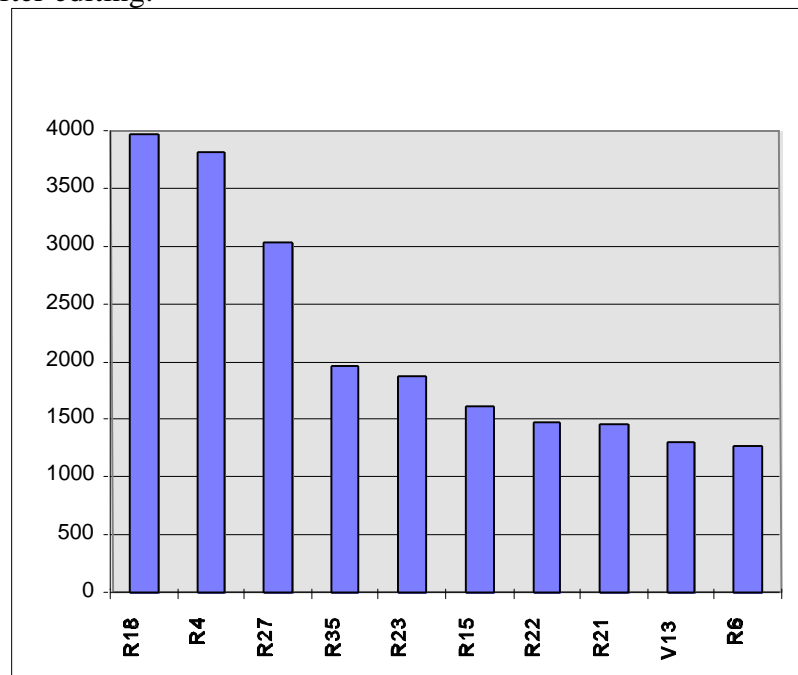


Figure 3: Number of flags by edit

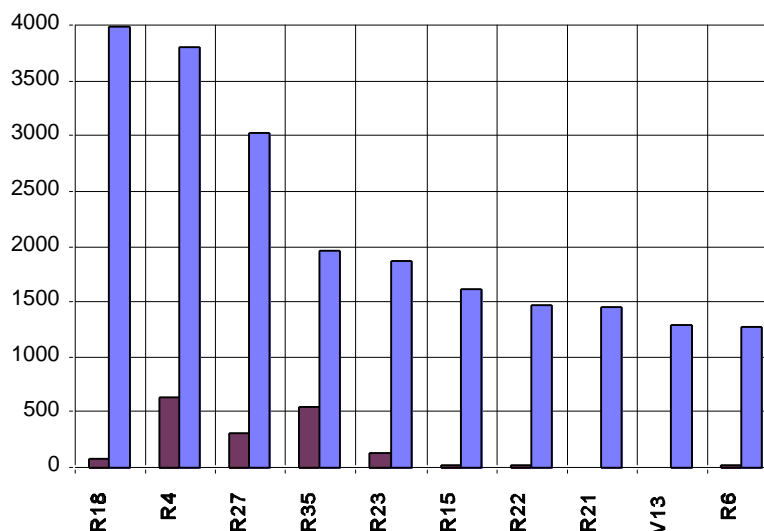


Figure 4: The number of flags before and after editing

The darker bar shows how many flags remain after editing.

13. Edit R18 is used for the question "For part-time work: give the agreed weekly working time in hours corresponding to full-time work". The edit checks if the given weekly working hours are 20-40 hours for teachers and 34-40 hours for other employees. Error cause data shows that 90 % of the flags are caused by respondent errors. A typical cause is that the weekly working time is not recalculated to correspond to a full-time worker. It is judged that most of these errors could have been avoided by changing the questionnaire and instructions to the respondent.

14. Generally, the edits are quite efficient. To reduce the editing work and the number of errors, the most crucial action seems to be to improve the questionnaire and the instructions to the respondent. Because the query edits in this survey are typically somewhat complex, with several variables involved, the indicators on variable level proposed in Engstrom(96) are difficult to achieve. However, if the error cause coding had been designed in such a way that it would be possible to detect which variables had been erroneous, it is judged that some very useful indicators at variable level could have been produced.

IV. CONCLUSIONS

15. The results of this study show that to keep the statistics on the data editing process is very useful, both to obtain information on edit performance and for evaluation of the questionnaire design. The indicators proposed in Engstrom(96) seem to be useful, but other data should be produced, for instance numerators and various forms of charts.

16. It is clear that the error cause data are extremely useful. The problem, though, is how to design a system which makes the coding less of a burden. One solution could be to only code a sample of the flagged units.

17. If a sample were used it would also be possible to code more thoroughly. It would be desirable, if possible, to see which variables were erroneous and also to obtain more specific information on why the respondent had made an error. A more thorough error cause coding could also give more extensive and accurate indicators and other data at variable level.

REFERENCES

Engstrom, P. (1996), Monitoring the editing process. *Conference of European statisticians, Work Session on Statistical Data Editing, working paper No. 9, Hague 1996.*