

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing
(Rome, Italy, 2-4 June 1999)

Topic (ii): Generalized software packages for statistical data editing, their evaluation

BLAISE III AND REDESIGNING THE FAMILY BUDGET SURVEY IN GERMANY

Submitted by the Federal Statistical Office of Germany¹

Contributed paper

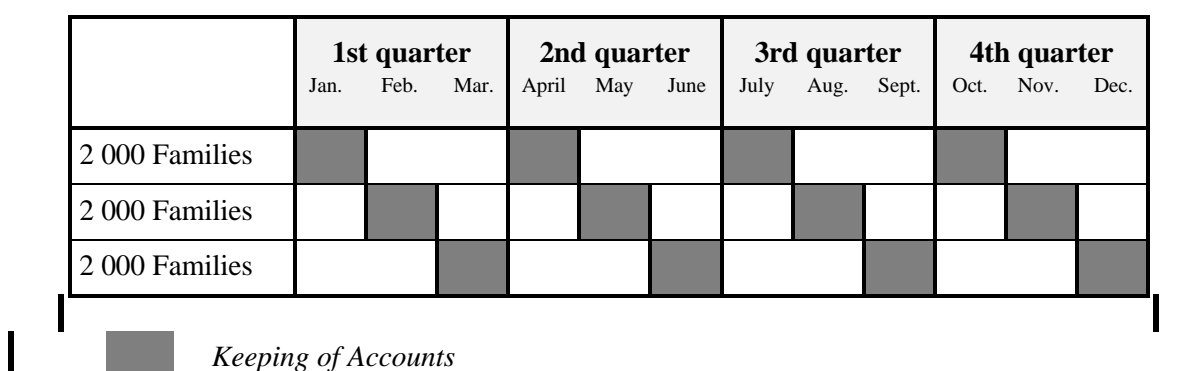
INTRODUCTION

1. This paper describes the demands made on Blaise III caused by redesigning the German Family Budget Survey (FBS). The experience gained during the development of an application with Blaise III and its subsequent testing at 5 statistical offices of the Länder of Germany are presented and suggestions for improvements of Blaise III are formulated. The use of Blaise III in the FBS of Germany will improve the timeliness of the statistical data, save costs, increase productivity and raise the motivation of statisticians.

I. DEMANDS ON BLAISE III CAUSED BY REDESIGNING THE FBS

2. The FBS is a survey performed monthly without obligation to provide information; the respondents are families - with the exception of employers, which is due to legal limitations. To reduce the burden on respondents, 2 000 families take part every month, while 6 000 different families per quarter are used for statistical results. This means that every family participates in only 4 different months a year. This fact leads to the following sample design:

Figure 1: Rotating Participation of Families in the FBS



3. All the families answer socio-demographic questions during a basic interview at the beginning of a year. They update parts of that information - for instance about employment - before indicating the receipts per person and the expenditures per day in a paper housekeeping account.²

¹ Prepared by Elmar Wein.

² For additional details of the survey design see Jürgen Chlumsky / Manfred Ehling (1997). Outline of the future scheme of household budget surveys. http://www.statistik.bund.de/mve/e/mad298_1.htm

4. The families are trained by 16 independent statistical offices of the Länder of Germany so that their entries meet the requirements of the classification of income and expenditure applied. The statistical offices collect the accounts every month, perform simple checks and then start data editing. Clean data of a quarter are transferred to the Federal Statistical Office of Germany (FSO) for aggregation at the federal level.

I.1 Specific demands made by the survey design on an EDP-system

5. Due to rotating participation, greater demands are made on survey management. Staff members of the statistical offices have to check whether a family takes part in the right month and correct data of the last family's participation have to be provided for data editing. More difficulties arise from the fact that some families wish to have a break over some months or want to change the month of participation. The expense for survey management should therefore be reduced by integrating a sample management in an electronic data processing system.

6. The sample size of the FBS is small but the survey contents are very voluminous and heterogeneous. Especially the housekeeping account represents high demands on an EDP system because of the underlying classification with nearly 400 headings and the predominant cardinal character of the data, which enables detailed comparisons between receipts and expenditures.

7. Some parameters which are needed for checks – such as the maximum rate of contributions to the national health insurance - change every year. The edp system should be adaptable according to the demands of a statistical office such as specific instructions or limits used for checks and imputations. To reduce the current edp support a lot of parameters should be stored in external files so that the users of the edp system can easily modify them.

8. Some of the main reasons for redesigning were the need for cost-saving and the improvement of timeliness. The intention, therefore, was to support all processes of statistical production such as raising, analysis, anonymising and calculation of standard errors and to reduce efforts for data transfer by integrating administrative components such as an address and payment management feature.³

I.2 Implementation of the survey design with Blaise III

9. The new survey design should be tested in a pilot study. The differences between the existing and the new survey design were so serious that a new EDP system had to be developed. The main reasons for using Blaise III were a very tight period for development (6 months), positive experience with Blaise 2.5 and a lack of experienced developers.

10. A CADI application was developed which allows a flexible statistical production in a network as well as on several stand alone PCs. About 46 Manipula⁴ and 4 Manipulus set-ups were implemented for a file management. It consisted of the administrative components mentioned above, a survey-specific help component, a simple training component, an interface used for data conversion from Blaise to ASCII and vice versa and an archiver (LHA) for data compression. Instead of a raising component, a weighting module compensating for families not included any more was developed in Manipula and a small analysis module based on Abacus was also integrated.

11. Nearly 283 survey characteristics were stored in 26 data models and about 838 checks were implemented. Twenty checks of one data model might be performed up to 500 times during one data entry

³ Finally all these considerations were summarized in an integrated data management. It is described by Hans Joachim Schwamb / Tatjana Theis / Elmar Wein (1998). Integriertes Erhebungsmanagement. In Methoden, Verfahren, Entwicklungen. Sonderausgabe 1/98, FSO, Wiesbaden, 1-3.

⁴ Manipula is a tool of Blaise III that is used for data manipulations. Manipulus is an extension of Manipula with interactive features used for the development of graphical user interfaces.

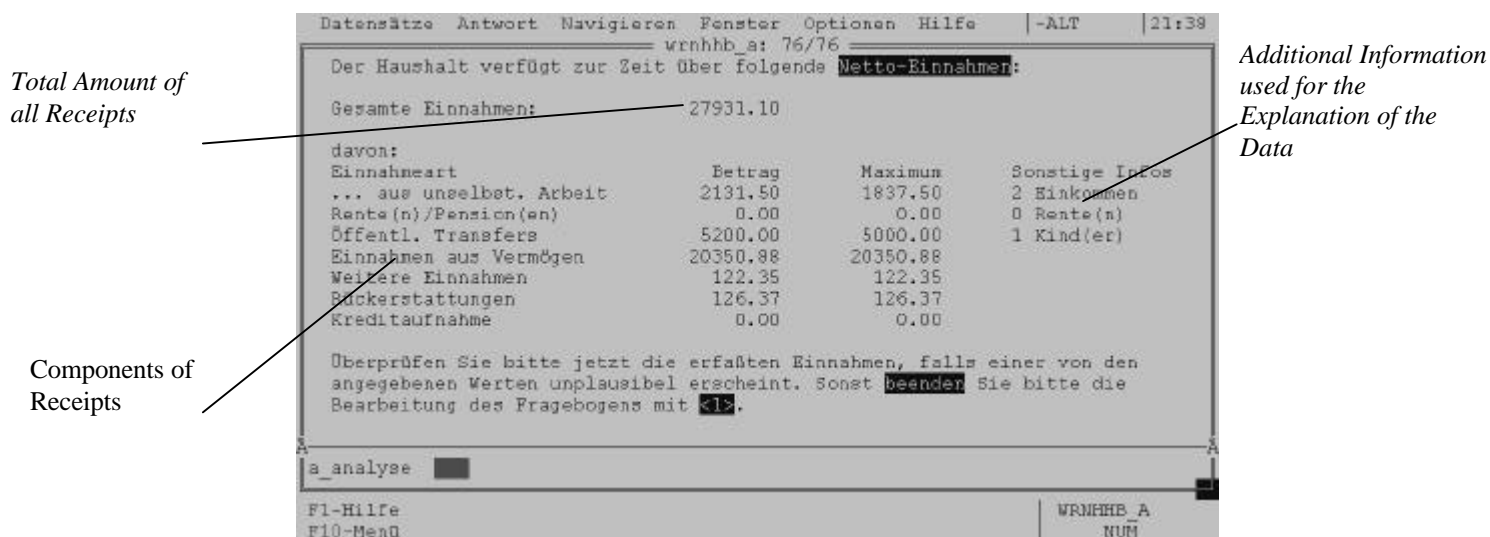
session. Checks were displayed in a standardized dialog box that contained an unequivocal check number, an error description, instructions for correction and a list of the involved fields.

Figure 2: Standardized Contents of a Check Message



12. The short time available for development did not allow the implementation of graphical editing components but one data model calculates 18 reference numbers from about 1,950 fields to obtain an overview of the family's economic situation.

Figure 3: A Screen giving an Overview of a Family's Economic Situation



13. One data model supports coding with 7 online classifications and two enlargeable libraries ("lookups") with nearly 13,000 predefined codes. Another one supports automatic coding. Range checking and important system parameters are stored in initial Blaise files that can be adapted by the users of the application.⁵

⁵ You can find an extensive description of the application by Hans Joachim Schwamb / Tatjana Theis / Elmar Wein (1998). *Integriertes Erhebungsmanagement*. In *Methoden, Verfahren, Entwicklungen*. Sonderausgabe 1/98, FSO, Wiesbaden, 4-7.

II. EXPERIENCE WITH BLAISE III

II.1 The developers' view

14. Blaise possesses very important general features:

- Its object-oriented approach offers powerful requirements of harmonising survey contents, checking and coding rules.
- The built-in opportunities of creating structures in relation to the selective checking mechanism enable the use of PCs even in the case of voluminous questionnaires.⁶
- The automatic generating of screens for electronic questionnaires, the mighty statistical functions such as lookups and hierarchical coding and the automatic adjustments provided in Manipula / Maniplus reduce the development effort and do not require detailed EDP knowledge.

15. Besides these advantages some experience was made during the development of the application. Only the most important experience concerning the general design of Blaise and its components are described in the following sections.

16. Blaise enables users to build voluminous and powerful survey management systems with lots of relationships between data models and Manipula/ Maniplus set-ups. The Blaise application of the FBS takes advantage of this fact but thereby causes a disadvantage in such a way that for instance a change in a data model used for sample management required the recompilation of up to 10 other data models and/or set-ups. A project explorer that automatically registers and indicates all informational relationships between different Blaise components would facilitate updating.⁷

17. The development of questionnaires and Manipula/Maniplus set-ups could be better supported. For example, the editor does not possess convenient functions like displaying the syntax even of self-defined modules during the writing of the code and the completion of an expression by offering a list of possible phrases. Due to the housekeeping account it was necessary to develop two data models; one of these electronic questionnaires produced a runtime code of 1.4 MB. The compilation, especially during the bug fixing, needed some time. It could be reduced with debugging tools like a permanent syntax checking, the opportunity to set breakpoints and a watch window.

18. For some checks of the receipts and expenditures, it was necessary to use the data of the families participating before. Such data are stored in different external files but on the basis of the same data model. Unfortunately paths of external files are not changeable in data models. The solution to this problem was the development of an additional data model and a Manipula set-up, which creates temporary files. An improvement would be the opportunity to change paths in data models by command line parameters.

19. One central feature of the application was that no additional lists should be used for statistical data editing but in some cases the users demanded them. Reports had to be created with Manipula by counting rows, columns and string lengths. An assistant for generating reports with the WYSIWYG feature would improve the layout and reduce the development effort.

20. The FBS contains a lot of cardinal data and may therefore lead to overediting. Blaise as a form-by-form oriented system offers a lot of specific statistical functions but no graphical tools for macro editing which enables an overview of the data structure in view of identifying outliers. Members of the Blaise

⁶ See the description of the selective checking mechanism by Mark Pierzchalla / Guus Razoux Schultz (1996). Optimal Instrument Performance in Blaise III. In Blaise Newsletter 8, International Blaise User Group, London, 11.

⁷ A project explorer is implemented in Blaise 4 which lists all related components of a project and enables a comfortable choice and loading of the modules but the functionality mentioned above is still missing.

programmers demonstrated test versions of graphical components but they have not been implemented yet in Blaise 4.⁸

21. Participation in the FBS places a burden on the families. Therefore, the issue of missing data should not be disregarded. One tendency of German statistics is the increased use of administrative registers on a long-term basis, which may also contain missing or completely wrong data so that the relevance of powerful imputation methods will increase. For the correction of missing data Blaise III offers only a simple hot deck imputation method based on the data of the previous questionnaire.

22. Data models with the Data Entry Program and Manipula/Maniplus set-ups form some crucial points of Blaise. A tabulation component should therefore preferably assist statistical data editing on macro level. In the FBS, information about receipts and expenditures is split into a lot of components that are presented only in a few tables to make interdependencies between them obvious. Abacus allows only 50 counter variables in one table so that the contents had to be split in many cases.

23. During the pilot study the users had only access to the runtime modules of abacus which produced predefined tables. The transfer of the tables into a worksheet software package like Excel was not possible. The conversion into a text writing programme completely destroyed the table layout produced by Abacus.

24. Clean data of the FBS were converted for further analysis into the format of SPSS 6.x. With Cameleon, Blaise III offers a tool to facilitate data transfer to other databases or statistical software packages. If you want to transfer data to SPSS 6.x for instance you should not choose field names longer than 8 digits in your data model because of limitations in this SPSS version. This aspect should be taken into account when developing a data model in Blaise. After relevant corrections had been made, the data transfer to SPSS did not pose any problems.⁹

25. Maniplus enables inexperienced programmers to develop complex applications on the basis of data models, Manipula set-ups and even external programmes. The Maniplus guide contains a lot of examples of Blaise applications but notes on the ergonomics of graphical user interfaces, the testing and implementation of complex Blaise applications are lacking.

26. Graphical objects like dialog boxes and a menu system can be developed with Maniplus by typing in the absolutely necessary information. Visual programming tools would provide a better support in these cases.

II.2. A test of the Blaise application at 5 statistical offices

27. In 1996, 5 statistical offices tested the Blaise application described above. At one office this was done by persons who normally are engaged in the FBS. A staff member of this office reported that using the application leads to a higher motivation of the personnel.

28. The application was run on stand alone PCs under different operating systems like different DOS versions, Windows 3.1 and Windows NT. In some cases, they caused an unexplainable application behaviour like corrupted files which had to be repaired with Monitor.¹⁰ The development team in the FSO collected all reports and developed a checklist for systematic testing of complex Blaise applications.

⁸ Graphical tools and a Manipula setup were presented at the International Blaise Users' Conference and described in the proceedings. See Jelke Bethlehem / Lon Hofman (1995). Macro Editing with Blaise III. In Essays of Blaise 1995 Proceedings of the Third International Blaise Users' Conference. Vesa Kuusela (Ed.), Helsinki, 5-21. Another non graphical procedure is described in detail in the Maniplus Guide. See CBS (1996). Maniplus Guide. Voorburg 1996, 155-171.

⁹ Problems of the data transfer caused by SAS are described by Fred Wensing (1997). Extraction of whole Blaise files to SAS. In Blaise Newsletter 9. International Blaise User Group, London, 13-15.

¹⁰ The use of another Blaise application in the sample survey of income and expenditure as a CAPI instrument shows that the system behaviour of Blaise on stand alone pcs causes less problems than its use in local area networks.

Performance tests in two different local area networks were done by the development team of the FSO and at one Land statistical office. The first results were that frequent access to central external files stored on a server slowed down data entry of the housekeeping account in spite of the read-only file attributes. Our impression is that in a network environment, it is harder to maintain a high performance speed with Blaise III than on stand alone PCs.

29. The application supports address and sample management and payment of the families. A lot of that information was already stored in Excel or Access files but could not be used by the application because there was no Open Database Connectivity feature of Blaise III.

30. In the course of the study it became clear that some of the hard checks had to be transformed into signals. These changes caused the recompilation of specific data model(s), dependent Manipula set-ups and the development of an update with a high degree of automatic processes. Attempts were made to reduce this effort. Tests at the FSO show that it is possible to store a flag about the kind of the check in an external file and to refer to it in a data model. This mechanism has not been tested in a survey but it seems that it would decrease the speed and extremely increase the size of the source code so that it might be a solution only for small data models. A mechanism similar to the check rules feature would reduce the maintenance effort in a very effective way. Blaise should be more flexible in this context.

31. In Germany, the decimal key on the number block of the keyboard makes data entry of numerical values impossible. Blaise III does not offer the opportunity to change the key code. We hope that this problem will be solved in Blaise 4.

32. Some members of the statistical offices asked for expressive error statistics but only a few functions are available for this purpose in Manipula. Especially modern quality management requires a protocol of quality changes to determine the reasons of quality defects and the processes which significantly contribute to quality improvements.¹¹ In view of data editing, original and changed values should be stored but Blaise only offers the questionnaire status or the history method which just indicate changes in values.¹² Another helpful feature will be a statistics about the number of (non) activated and suppressed checks or rather signals which would supply valuable information about the check design and/or the appropriation of the questionnaire.

33. The above statements represent, first of all, the desire to improve an effective and powerful system. Problems mentioned above could have been reduced with more experience and some of them are regarded as negligible. On the whole, the test of the Blaise application was passed successfully. Staff members of the statistical offices expect that the application:

- will permit saving money for data entry of up to 62 000 Euro per statistical office per year,
- will reduce the time required for obtaining the quarterly results by 1 - 2 weeks and
- will increase average productivity per person from 25 housekeeping accounts per month to 40.

Beyond these expectations comparisons with similar surveys like the five-yearly sample survey of income and expenditure gave no significant hints of a deterioration of data consistency and comparability.¹³

34. The positive results of the pilot study facilitated the permission of the redesign by the conference of the heads of all statistical offices in 1997 so that the survey started with a modified design and a Blaise application in January 1999.

¹¹ See for instance Deutsches Institut für Normung (1994). Quality management and quality system elements - Part 1: Guidelines (EN ISO 9004-1). Berlin, 33.

¹² CBS (1996). Blaise Reference Manual. Voorburg 1996, 173.

¹³ For additional information about the statistical results see: Felix Gertkemper / Carola Kühnen / Elmar Wein (1998). Ergebnisbericht der Testerhebung zur Neukonzeption der Laufenden Wirtschaftsrechnungen. FSO, Wiesbaden, 35-55.

III. SUMMARY

35. The test of a Blaise application during the redesign of the FBS showed that Blaise III is a survey management system with:

- well designed rough drafts,
- practicable automatic adjustments and
- powerful functions for statistical data editing

which:

- decreases the development effort,
- improves productivity and timeliness of statistical data and
- enables powerful and voluminous questionnaires by
- relatively low demands on stand alone PCs.

36. The advantages still exceed very clearly the deficiencies of some components but there is a growing importance of graphical analysis tools, powerful imputation methods and features for the support of a quality management.

37. Based on the good experience acquired with Blaise in several German population surveys, the FSO and 16 statistical offices of the Länder plan to expand their Blaise engagement by becoming corporate software licence users in 1999.