Automatic Generation of Blaise Data Models

Saliha Zayoum and Lars Peter Jørgensen, Interview Service division, Statistics Denmark
Leif Bochis Madsen, IT Center, Statistics Denmark

Abstract

The survey division of Statistic Denmark has utilized a system to speed up the process of authoring Blaise questionnaires.

The system is a compromise between on the one hand, an easily adoptable way of describing the questionnaire for customers and methodologists unaware of Blaise syntax and, on the other hand, a document format that is manageable for automatic processing.

Customers and methodologists can describe questions, answers and filters grouped in tables in a MS Word document, questionnaire designers can add information about web form layout etc. and the system may automatically produce a Blaise data model almost ready for testing and deployment.

The system has proved efficient as it has reduced the time needed for production of Blaise questionnaires considerably.

1. Introduction

The Interview Service division (IS) is a central unit in Statistics Denmark responsible for collection of data via telephone interviewing and via combined telephone and web interviewing. Interview tasks may be requested by customers – external as well as internal (within our organization). Our external customers include public institutions, universities, organizations, private companies and EU.

The primary purpose of using auto generation is to standardize and optimize our procedures. The auto generation results in a faster set-up of Blaise data models for the use of combined CATI/CAWI surveys. As a basis of further processing an MS Word document is produced – structured mainly with the aim of auto generation in mind.

Another important purpose is to involve the customer in the detailed specification of the questionnaire and to have the costumer make sure that the content of the survey is correct and accepted before the set-up. For this purpose, the produced document also serves as a reference document as part of the agreement between customer and IS.

The different types of customers participate in the development of questionnaires at different levels of involvement. Typically, EU delivers a questionnaire which is already final and cannot be changed. In these situations we copy the received questionnaire into the word document as a preparation of the Blaise set-up.

Other customers deliver a non-finished questionnaire which we can comment on and make corrections. The development process is a matter of cooperation between the customer and survey experts at Statistics Denmark.

1.1 Why Word?

The format of the reference document has evolved over some years of experiments with auto generation using various formats of Excel spreadsheets and Word documents. We decided to use Word documents instead of using e.g. Excel spreadsheets, mainly because of the widespread usage and because we wanted a tool that was simple enough to expect our customers to use it without too much instruction. Word is a well-known tool and some of our customers are not as familiar with spreadsheets as they are with word processing. Furthermore it is possible to specify e.g. formatted text to be inserted in the electronic questionnaire automatically. In spreadsheet documents there may be limitations on the size and formatting of text.[2] There are no practical limits on the amount of space using MS Word and the set of constructions we use in the documents are widely supported also by any other word processing tool. In the end it is a compromise between different purposes, where ease of use is considered most important.

1.2 Flow of document

Users of the reference documents are:

1) Customers
2) Survey and questionnaire designers (IS)
3) Blaise questionnaire developers (IS)

The document is used by different users who have the opportunity to contribute to the design of the questionnaire/survey. The different kinds of users, however, typically contribute with different levels of detail.

Our customers are limited to only describe the content of the questionnaire. They are also able to note their wishes of the web-design of the survey. E.g. if they want a special set-up as a group table or a matrix they can make a note of that.

We receive a non-finished questionnaire from the customer which our survey designer critically comments on. Afterwards the customer receives the document with the purpose of accepting the questionnaire.

The document is usually sent back and forth with comments and corrections between our internal survey designer and the customer and sometimes the Blaise programmers before the set-up in Blaise.

Our customers usually do not have any Blaise experience and it makes it difficult for them to describe the routing of the questionnaire. In those cases we construct the routing on their behalf to make it easier.

The Blaise programmers may add/correct technical details before using it for auto generation. Of course, a part of this role is to make sure that the field names and the logic (filters etc.) are correctly defined.

[2] Recent versions of MS Excel allow larger texts as well as formatting (bold, e.g.). However, it is not attractive to depend on the products and versions of the office packages our cutomers use.

2. Description of the document

The document must be structured in accordance with the following definition:

The document consists of tables. Each table describes a section (block) of questions (fields) in the questionnaire (datamodel). The tables consist of rows and columns.

The tables are structured so the first row describes the block and the other rows describe the fields in the block.

2.1 Table

Fundamentally a table is divided into 6 columns:

1. Fieldname
2. Text
3. Type name
4. Type definition
5. Filter (condition)
6. Comments

Example: A table defining a block with one question

BlockFieldName				Filter (optional)	Remarks
FieldName	Question Text	TypeName	Type Definition	Filter (optional)	Remarks

2.2 Definition of Blocks

The first line in the table describes the block and the content of the cells will result in the following generation:

```
BLOCK B<BlockFieldname> ”<Text>”
FIELDS
   <…description of the fields…>
ENDBLOCK

FIELDS
   <BlockFieldname> : B<BlockFieldname>

RULES
   IF <Filter> THEN <BlockFieldname> ENDIF
```

Or if no filter is defined:

```
RULES
   <BlockFieldname>
```

The text in the second column is often omitted for block definitions. Type name (third column) and type definition (fourth column) are irrelevant regarding block fields. The type name is always produced as a concatenation of “B” and the given field name. It is possible to denote a keyword “TYPE” at the second line of the first column in order to identify blocks that are used more than once. These blocks may then be

referenced in the type name column of field rows in other tables. Certain layout instructions may occur in the third column.

2.3 Definition of Fields

Generation of field definitions follows a different pattern. Rules are generated in the same way as blocks but fields are generated as follows:

```
FIELDS
   <Fieldname> ”<Text>” : <Type name>
```

If a type definition is given the following will be generated:

```
TYPE
   <Type name> = (<Type definition>)
```

The type can be enumerated as well as a numeric interval.

2.3.1 The first column (fieldname)

The first section (line) in the cell is the fieldname. The content must follow the rules for field names. Typically it consist of letters and numbers e.g. A1, A1_1a and so forth. Blaise cannot accept numbers alone as a fieldname or a fieldname starting with a number. Our customers usually assign the fieldnames and we only make changes if necessary.

The additional lines are ignored, but later on they can be used to define different versions of the same question, e.g. a version only for CAWI/CATI or versions of different languages besides Danish.

2.3.2 The second column (text)

The text is transformed into question text in Blaise syntax. Empty lines between the texts are transformed as line breaks.

Italics, bold, underline are transformed into Blaise text code (@B, e.g.).

In the second column we and our customers are able to define the final set-up of the text, line breaks etc.

2.3.3 The third column (type name)

The first section (line) is transformed into type name and is used directly in the field definition. If you have a question with the option Yes/No (YN) it means that YN is a type name. Instead of a type name it is also possible to indicate other type statements such as STRING, STRING[250], 1..100, Datetype, Open etc.

Furthermore a comma will result in an end of the type name. E.g. you can define a type name as YN, DONTKNOW; YN, REFUSAL or YN, EMPTY. The type name will be YN and the rest will be used as an attribute to the field.

The second section (line) can be used to modify the type name with e.g. a SET OF or ARRAY[1..10] OF. This modification will be placed before the type name in the field definition. Also, it must be in

compliance with the Blaise syntax. If there is content in the third section (line) it will be used as an instruction to the layout for CAWI. Note that it is the name of the field pane which can be noted in the third section.

An example of the content in the third column:

```
MyAnswerType, EMPTY
SET [3] OF
```

”MyAnswerType” is the field type and the field definition will be produced as:

```
SET [3] OF MyAnswerType, EMPTY
```

Another example:

```
MyAnswerType, DONTKNOW

DropDown
```

The field will have the type name MyAnswerType and respondents may answer the question with DONTKNOW. Furthermore a layout instruction “At fieldname FIELDPANE DropDown” will be created. Notice that the second line is empty. As mentioned the second line is reserved for SET OF and/or ARRAYs.

2.3.4 The fourth column (type definition)

The type definition can be omitted, but it will typically be used in combination with the third column to define enumerated types or numerical intervals which will be re-used.

During the definition of enumerated types each section (line) will be used as an answer. Empty sections (lines) will be ignored.

If a numerical value is defined in a bracket in the beginning of the line e.g. “(0) None”, the answer category will be assigned the given code zero in the type definition.

An example of the content in the fourth column:

```
None
1-4
5-10
```

If a type name is assigned in the third column e.g. “MyAnswerType”, a type will be defined as:

```
MyAnswerType =
        (s1_None ”None”,
         s2_1_4 ”1-4”,
         s3_5_10 ”5-10”)
```

And the answers will be given the code values 1, 2 and 3.
Another example:

```
None
(5) 1-4
5-10
```

If a type name is assigned to ”MyAnswerType” in the third column, a type will be defined as:

```
MyAnswerType =
        (s1_None ”None”,
         s5_1_4 (5) ”1-4”,
         s6_5_10 ”5-10”)
```

And the answers will be given the code value 1, 5 and 6.
Notice that any automatic numbering of answers in the word document will be ignored e.g.:

0. None
1. 1-4
2. 5-10
3.

And will be transformed as shown in the first example.

2.3.5 The fifth column (filter)

A filter defines a condition whether a question should be asked or not. A filter can be defined in a simple syntax where code names and code values can be used, e.g.:

a) Question5 = Yes
b) Question5 = 1
c) Question6 = 1-4
d) Question6 = 1-2, 5, 7-9

If we presume that Question5 is defined with the field type name YN, then (a) and (b) will be interpreted in the same way. Again, if we presume that Question6 is an enumerated field with at least 9 answer categories, then (c) and (d) will be interpreted as answer quantities and transformed into:

c) IF Question6 IN [answer1..answer4] THEN
d) IF Question6 IN [answer1.. answer 2, answer 5, answer 7.. answer 9] THEN

If you have a filter for the whole block, it can be defined in the first row of the table e.g.:

Example: Definition of a block with an introduction and a question

BlockB	Discrimination			Citizenship <> Danish	The questions in BlockB are not asked of Danish citizens
IntroB	The following questions deal with various types of discrimination you may have experienced in Denmark due to your ethnic background				
B1	Within the past year, have you experienced, because of your ethnic background, being denied access to places which other people were allowed to enter? (such as a bus, taxi, nightclub or swimming baths)	YNDR	1 Yes 2 No 3 Do not know 4 Prefer not to answer		

Provided the background information includes citizenship of the respondents we can structure a filter for the entire block as defined in the table above. This means that only the non-Danish respondents are asked the questions in BlockB.

2.3.6 The sixth column (comments)

The column is used for remarks to the field. The remarks are inserted as comments immediately after the fieldname in the RULES section of the block. E.g. remarks about signals and checks. Comments may be added in free format.

Also, it is possible to assign a field description in this column by using s special syntax:

DESCRIPTION *my description…* ENDDESCRIPTION

3. Technical description of the generation process

In the following we will describe the single procedures in the process of generating a Blaise web questionnaire from a description in a Word document.

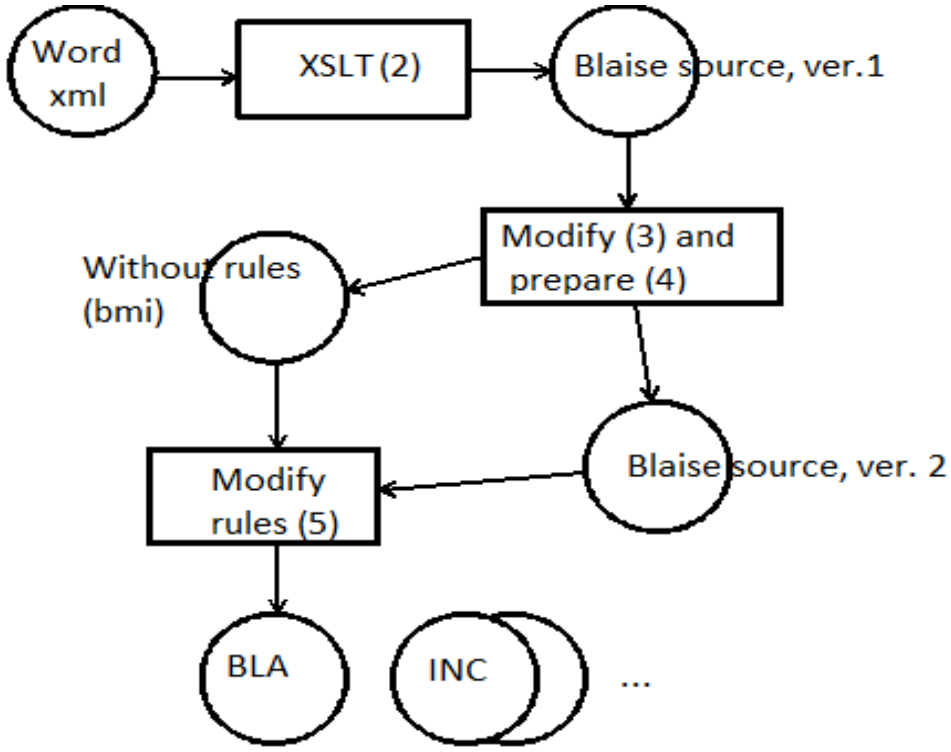
The main instrument for the generator is an XSL program transforming the Word document (in XML format) into a text file while producing Blaise syntax. This program is supplemented by a number of Manipula programs and VB scripts.

Steps of generating Blaise source code from a word document comprise:

1) Conversion of word document into an xml representation
2) Transformation of xml document into Blaise source code
3) Modification of Blaise source code
4) Preparation of a Blaise data model without rules instructions
5) Modification of rules instructions
6) Inclusion of generated source code into basic template
7) Preparation of final Blaise data model
8) Generation of layout groupings
9) Generation of a Blaise database (test data)
10) Validation of Blaise Internet Specification (.bis)
11) Generation of a Blaise Internet Package (.bip)

Generation is an iterative process: At any step the generation may be halted. For example, if the Word document deviates from the requirements, the generation may be halted and the Word document has to be corrected before regeneration.

Fig.: Steps 2-5 of the transformation proces



3.1 Conversion of word document into an xml representation

Initially, the word document is converted into xml representation (Word 2010 format) in order to facilitate the use of XSL transformations.

Xml format is a complete representation of the word document – actually, the Word 2010 file format (.docx) is merely a compressed file containing xml. Therefore, conversions from docx to xml or vice versa can be done without loss of data. The task is automated using a VB script referencing the Word API.

3.2 Transformation of xml document into Blaise source code

XSL transformations comprise identification of the constructs of the tables, rows and columns in the Word document. Each table is converted into a block definition consisting of:

1) The block definition itself
2) A field definition defining an instance of the block (unless it is marked TYPE)
3) An ASK instruction for the rules part of the main questionnaire block (unless it is marked TYPE)

From each of the rows in the table, proper FIELDS, RULES and LAYOUT instructions are generated with respect to the specific properties of the different columns. For example, the second column is used for definition of question text and requires that formatting codes – e.g. italics and bold – are preserved and converted into Blaise syntax. The fourth column may contain answer texts which should be handled the

same way, while all other columns contain only syntactical definitions where any formatting codes must be ignored.

3.3 Modification of Blaise source code

The source code generated by XSL transformation needs some modifications. For example, a Word document may usually contain characters that are not representable in the Windows character set used for Blaise source code.[3] Therefore, the XSLT process produces output in utf-8 which will need to be converted.

Also, the XSLT process stores the generated source code in one single text file. We will prefer to split this file into several files – one for each defined block.

These conversions are handled by a Manipula program.

3.4 Preparation of a Blaise data model without rules instructions

The most difficult part of the source code generation is the proper generation of Blaise rules instructions. Therefore, the previous process also generates a Blaise data model without rules instructions. The first reason is that at this stage we can try to prepare a Blaise data model without rules and layout instructions, but comprising the overall structure including blocks, fields and types.

If the preparation of this data model fails then there is something wrong with the specification in the Word document and there is no reason to proceed any further. However, if the preparation is successful we may use this data model for looking up fields and types and retrieving complete path names for the fields and checking or converting between code names and codes of enumerated fields.

3.5 Modification of rules instructions

Conditions and expressions need to be properly checked and this is done by a Manipula program using the metadata functions (GETFIELDINFO, mainly) to look up the different elements of an expression in the rules-free data model prepared in the previous step.

The rules parts of the source code are searched for structures like “IF condition THEN” and for lines containing “name := something”. By looking up the fields and their types in the data model names of fields and categories may be checked, constants may be checked for validity and – possibly – field paths may be added to fields belonging to another block than the block currently analyzed.

3.6 Inclusion of generated source code into basic template

The basic Blaise-based data collection depends on usage of common templates for different kinds of data models.

Business and household surveys, for example, differ in the kind of background information needed and specific CATI handling. However, most of the differences may be defined in a few blocks holding the

[3] The generator was initially made to produce code for Blaise 4.8.2. From Blaise 4.8.3 it is possible to represent question texts etc. as utf-8, but we haven’t experimented with this representation.

background data and non-response codes and treatment. Also, there are differences in the way web and telephone surveys are carried out. Therefore, we use separate data models for the two modes, but these data models only differ in the general handling of the form and still share the same questionnaire.[4]

The general template contains hooks that can be used for automatic inclusion of the generated code.

For example, in the main questionnaire block the template contains a hook:

```
{### INCLUDES BEGIN ###}
```

At which place the following code could be inserted:

```
INCLUDE "BBlockA.inc
INCLUDE "BBlockB.inc"
INCLUDE "BBlockC.inc"
```

Likewise, there are defined hooks for placement of FIELDS and RULES instructions for the main questionnaire block.

This task is also carried out by a Manipula program.

3.7 Preparation of final Blaise data model

If the previous steps were successful, the Blaise parser will prepare the data model!

3.8 Generation of layout groupings

If the preparation was successful, it is possible to inspect the data model in order to generate layout groupings for the web questionnaire. The layout generator is a C# program using the Blaise APIs in order to retrieve information about field pane settings from the generated data model and to generate and store layout groupings in the Blaise Internet Specification file.

There are a number of conventions applied in order to figure out which groupings should be generated. For example, a series of fields on the same page all using the same GroupTable field pane will result in the generation of a GroupTable layout group comprising all these fields.

Also, the mere existence of two or more consecutive fields with the first field of type enumeration or set and the second and possibly following fields of type string or numeric and using the field panes named OtherSpecify will result in the generation of an Other-Specify group.

[4] The generator is included in the Cati Survey Management System, maintained since 2000 and described in a paper to the IBUC/2003. Templates and generators have been an integral part of this system for many years.

3.9 Generation of a Blaise database (test data)

For the purpose of testing the questionnaire a test database should now be initialized. Depending on the needed background information the data set can be a standard test data set or a data set created for the specific survey modified for the purpose of testing.[5]

3.10 Validation of Blaise Internet Specification (.bis)

Through a call to the IsValid method of the Blaise Internet API the modified specification file may now be checked for validity. If it is valid, the generation process has been completed and the questionnaire is ready for testing. This task is carried out by a VB script.

3.11 Generation of a Blaise Internet Package (.bip)

Last step of the automatic generation is construction of a package file (the same VB script as above), possibly followed by installation on a web server (by another VB script referencing the Blaise Internet Server API).

4. Conclusions

The automatic generator has been in use since the beginning of 2012 and has proved useful in order to:

- Save time in the production of questionnaires
- Produce a source code more safely and with fewer errors
- Improve the cooperation with the customer in the process

The time saved is mainly in the more "boring" parts of questionnaire generation which allows the questionnaire developers to focus their efforts on the more difficult parts of questionnaire development, i.e. the rules and conditions, and also to improve the layout of web questionnaires.
Therefore, automatic generation has led to an improvement of the quality of our questionnaires within the given and unchangeable time frames the development process is subject to.

Development of the generator has been focused on current needs, i.e. the requirements have been formulated stepwise as the needs have turned up. Thus, the features developed have been the most important and work-saving.

A number of enhancements are still on the list and may be implemented as they reach sufficient importance. For example, we can mention support for randomization procedures and there is still discussion going on how we should supply more support for the data delivery, e.g. by incorporating descriptions in the document.

[5] Test data generators and standard input programs has been part of the SMS for a number of years. They have been implemented using Manipula and C#.

5. Acknowledgements

We wish to thank the Blaise community for long lasting and varied inspiration in how to do as much as possible with as little effort as possible. Special thanks to Gerrit de Bolster, Statistics Netherlands, who kindly submitted manuals and descriptions of the Blaise IS generator. All these different ways to generate and get the job done has been a great inspiration for the development of our own generator.

6. References

Gerrit de Bolster: BlaiseIS at Statistics Netherlands, in: Essays on Blaise 2009. Proceedings of the 12th International Blaise Users' Conference, Riga 2009.

Appendix: Questionnaire "2012 Integration Barometer"

BlokA	Citizenship				
IntroA	First, a few questions about your participation in Danish society				
A1_2a	Denmark has many associations and clubs, such as trade unions, sports clubs, tenants' associations, cultural and religious associations, consumer societies like Brugsen and FDM, as well as fundraising organizations like the Red Cross and the Danish Cancer Society. Are you a member of a club or an association?	YNDR	5 Yes 6 No 7 Do not know 8 Prefer not to answer		
A1_2b	Within the past year, have you taken part in a meeting or other activities held by an association or club of which you are a member?	YNDR		A1_2a = 1	
A1_2c	Within the past year, have you carried out unpaid, voluntary work for any association or club of which you are a member?	YNDR		A1_2a = 1	
IntroA1_3	The following questions deal with your political participation in society	QuestionTextOnlyGrouptable NEWPAGE			
A1_3a	Within the past year, have you been a member of a political party or taken part in a political meeting aimed at social change?	YNDR GroupTable			
A1_3b	Within the past year, have you written a letter to the editor of a newspaper or taken part in a debate on the internet aimed at social change?	YNDR GroupTable			
A1_3c	Within the past year, have you taken part in a petition drive, a demonstration or a strike aimed at social change?	YNDR GroupTable			
A1_3d	Within the past year, have you contacted a politician, a public official, the media, an association or an organization for the purpose of social change?	YNDR GroupTable			
A1_3e	Within the past year, have you boycotted or deliberately chosen to buy specific products, such as organic products, for the purpose of social change?	YNDR GroupTable			

A1_3f	Within the past year, have you been involved in fundraising for or contributed to an organization like the Danish Cancer Society or a political party for the purpose of social change?	YNDR GroupTable			

BlkB3MatrixRow TYPE		NONEWPAGE			
B3_YN	Yes/No	YNDR_DD MatrixFieldpaneDropdown	Yes No Do not know Prefer not to answer		DESCRIPTION Yes / No: ENDDESCRIPTION
B3_HowM	Within the past year, how many times?	TManyTimes_DD MatrixFieldpaneDropdown	1 Once 2 2 – 5 times 3 6 – 10 times 4 More than 10 times 5 Never 6 Do not know 7 Prefer not to answer	B3_YN = 1	DESCRIPTION Within the past year, how many times? ENDDESCRIPTION
B3_Where	Where (most often)?	TPlace_DD MatrixFieldpaneDropdown		B3_HowM = 1-4	DESCRIPTION Where (most often)? ENDDESCRIPTION

BlokB	Discrimination			UndersType 1 = 1,2	The questions in Block B are not asked of Danish citizens
IntroB	The following questions deal with various types of discrimination you may have experienced in Denmark due to your ethnic background				
B1	Within the past year, have you experienced, because of your ethnic background, being denied access to places which other people were allowed to enter? (such as a bus, taxi, nightclub or swimming baths)	YNDR			
B1_1	*Within the past year*, how many times have you experienced,	TManyTimes	8 Once	B1 = 1	

	because of your ethnic background, being denied access to places which other people were allowed to enter? (such as a bus, taxi, nightclub or swimming baths)		9 2 – 5 times 10 6 – 10 times 11 More than 10 times 12 Never 13 Do not know 14 Prefer not to answer		
B1_2	Where did you experience this most often?	TPlace_DD Dropdown	1 In the street 2 On a bus, train, taxi, airplane or the like 3 In a shop 4 In a bank 5 At a café, bar, nightclub or the like 6 At swimming baths, a water park, an amusement park or the like 7 In a sports club or a cultural or religious association or the like 8 At or near your home 9 At work 10 While attending school, training courses or the like 11 In contact with a doctor, nurse, hospital or the like 12 In contact with the municipality 13 In contact with the police 14 At a court of law 15 Other 16 Do not know 17 Prefer not to answer	B2 =1-4	Dropdown
B2	Within the *past* year, have you experienced, because of your ethnic background, being rejected after having applied for a job, bank loan, housing, mobile phone subscription or similar?	YNDR NEWPAGE			
B2_1	*Within the past year*, how many times have you had an application rejected because of your ethnic background?	TManyTimes		B2 = 1	
B2_2	Where did you experience this? (more than one answer is possible)	TPlaceVar SET OF	1 In the street 2 On a bus, train, taxi, airplane or the like 3 In a shop 4 In a bank	B2_1 = 1-4	Validering: Ved ikke kan ikke kombineres med andre svar.

			5 At a café, bar, nightclub or the like 6 At swimming baths, a waterpark, an amusement park or the like 7 In a sports club or a cultural or religious association or the like 8 At or near your home 9 At work 10 While attending school, training courses or the like 11 In contact with a doctor, nurse, hospital or the like 12 In contact with the municipality 13 In contact with the police 14 At a court of law 15 Other 16 Do not know 17 Prefer not to answer		SET OF
B2_2Other	Where?	STRING[100] InputAndErrorText		B2_2 = 15	Other specify

BlokB3Matrix				UndersType 1 = 1,2	
B3Intro	*Within the past year,* have you experienced the following situations because of your ethnic background? (such as in a shop, at a café or hospital, at the doctor's, from the municipality or from the police)				Matrix
B3_1	Being given poor service?	BBlkB3MatrixRow			DESCRIPTION Being given poor service? ENDDESCRIPTION
B4	Being the object of offensive words or degrading jokes?	BBlkB3MatrixRow			DESCRIPTION Being the object of offensive words or degrading jokes? ENDDESCRIPTION
B5	Being spat on, shoved, jostled or hit?	BBlkB3MatrixRow			DESCRIPTION Being spat on, shoved, jostled or hit? ENDDESCRIPTION

BlokC	Social supervision				
IntroC3	The following questions deal with your family's role in your choice of boyfriend/girlfriend and spouse			Alder = 18-29	
C3_1	Are you married?	YNDR		Alder = 18-29	
C3_1a	Did your family permit you, or do you think they permitted you, to have a boyfriend or girlfriend before you were married?	YNDR		C3_1 = 1 AND Alder = 18-29	
C3_1b	To what extent do you feel that your family has let you freely choose your present spouse?	TExtent	To a great extent To some extent To a lesser extent Not at all Do not know Prefer not to answer	C3_1 = 1 AND Alder = 18-29	
C3_1c	Did your family choose a spouse for you against your will?	YNDR		C3_1 = 1 AND C3_1b= 3, 4, 5 AND Alder = 18-29	