The internet, Blaise and a representative panel

Adriaan Hoogendoorn
Dirk Sikkel
Bas Weerman

30 march 2000

The internet, Blaise and a representative panel

Abstract

For 10 years CentERdata has operated a household panel that is representative for the population in the Netherlands. Recently, the production system has switched to the internet, using Blaise software. It is described how the production system works and how Blaise fits in. Adaptions had to be made to make complex page layout possible when a respondent uses a relatively simple settop box. Two examples of complex surveys are given. The first example is a large and complex financial survey that is repeated every year. The second example is a survey on product innovations which is repeated monthly.

1. Introduction

Collection of household data with electronic means has been a major development during the last 20 years. The most common tool so far is the laptop computer. Interviewers use them to enter the data in the presence of the respondent thus enbling the respondents to correct inconsistencies and statistical agencies to produce reports immediately after data collection. These CAPI (Computer Assisted Personal Interviewing) procedures nowadays are used by many governmental statistical agencies and commercial market reserach firms. With the rise of the internet, however, new ways of household data collection seem to arise: the internet interview. Respondents fill in questionnaires they find on the internet, and mail the data to the address of the data collector. With respect to this practice, several observations can be made.

1. The idea of direct electronic household data collection is not new. CentERdata has carried out such interviews already for more than 10 years, using its own communication channel. The data were collected in a 'telepanel', a panel of households who have a PC (if they don't own a PC, they may borrow from CentERdata). The households fill in a questionnaire every week.

2. Data collection over the internet takes place without interviewers. This makes the design of the questionnaire even more critical than in the case of CAPI. Extra care has to be taken that ambiguity is avoided, that the respondent has all the information he needs (a good help function), that no routing errors are made and that the respondents are motivated to give valid answers.

3. Interviewing over the internet raises the question of representativity. When a questionnaire is placed on a fixed location, waiting for respondents to pass by, there is no way of knowing what the composition of the set of respondents is. The same is true when a questionnaire is sent out to a random group of internet users. Even worse results may be obtained when from a large group of internet users a group is selected that is interested to take part in a survey. The participants to such a survey may have a specific interest in its outcomes.

4. The anarchistic character of the internet makes it possible for virtually everybody to do research on the internet. Household data collection is, however, a discipline that requires talent, training and skills. It is important that it becomes clear which firms have the necessary skills and means to succesfully execute research projects using the internet.

In this paper, we address the aspects which determine the quality of the collected data. The first aspect, selection of respondents, is discussed in section 2. Section 3 deals with executing and monitoring the production process. The technical execution of the interviewing process is the subject of section 4. The data quality is, of course, also determined by how the questionnaire is designed from a substantive point of view. Two examples are given in section 5. Section 6 concludes.

1. A controlled representative panel

CentERdata selects its panel, the CentERpanel, in such a way that the panel is a good cross section of the Dutch population. The selection procedure takes place in two steps.

1. *The recruitment interview*. This is a CATI interview among a random sample of the Dutch population. This is a short interview. Its main purpose is to determine if households are willing to take part in the CentERpanel. Respondents who do not refuse to consider taking part in the panel are moved to a database. In this database their demographic characteristics are stored. In the CATI interview the respondents are questioned, apart from demographics, about a variety of subjects including health, work, victimization and participation to cultural activities. The resulting data, including those which belong to respondents that refuse to partipate in the panel, are used as reference distributions for weighting panel members.

2. *The selection interview*. New potential panel members are selected from the data base. This selection is based on demographic characteristics. Those potential members are selected who bring the distributions of age, income, composition of the household, region, urbanisation and voting behaviour as close as possible to the corresponding population distributions.

In this process, there are three moments on which persons may refuse to participate. In the first place, a person may refuse to participate in the recruitment interview; 62% of the persons which are selected take part in the recruitment interview. At the end of the recruitment interview, respondents may indicate that they refuse to participate in the panel; 52% is willing to consider participation in the panel. Finally, when the respondents are asked to actually become a panel member, 35% of the respondents agrees. Thus, the participation rate is 0.62*0.52*35% = 11%. This may seem a small number, but with respect to the variables mentioned in 2, the panel still is forced to be representative. Moreover, we can compare distributions of the variables in the recruitment interview and the selection interview to the actual panel data. Table 1 shows that the distributions with respect to health and cultural participation are somewhat dissimilar. Of course, by weighting the data, this dissimilarity can be adjusted.

Table 1. Distributions of variables in the recuitment interview, the selection interview and in the panel

	Recruitment %	Selection %	Panel %
buy/rent dwelling			
Buy	72.8	74.1	68.7
rent	25.8	25.2	30.1
other	1.4	0.7	1.2
number of rooms in the house			
1-3 rooms	13.1	12.1	17.7
4 rooms	37.3	42.9	31.8
5 rooms	28.3	27.3	30.9
6 or more rooms	21.3	17.6	19.5
traveling time to work			
>20 minutes	40.7	52.0	49.3
<20 minutes	59.3	46.8	50.7
ill at home last 3 months			
Yes	20.9	21.3	27.0
No	79.1	78.7	73.0
chronic disease			
Yes	18.4	21.1	22.5
No	81.6	78.9	77.5
rating personal health			
1 through 5	6.0	2.9	7.3
6	6.3	5.9	10.8
7	22.6	21.7	27.5
8	40.0	46.9	36.3
9, 10	23.8	22.6	18.1
visited the cinema last year			
Yes	48.6	53.5	38.4
No	51.4	46.5	61.6
No	51.4	46.5	61.6
visited the theater last year			
Yes	31.7	38.7	45.3
Yes	31.7	38.7	45.3
No	68.3	61.3	54.7
No	68.3	61.3	54.7
victim of a burglary			
Yes	16.4	18.4	22.8
Yes	16.4	18.4	22.8
No	83.6	81.2	77.2
No	83.6	81.2	77.2
sometimes afraid at home			
Yes	8.2	8.0	7.5
Yes	8.2	8.0	7.5
No	91.8	92.0	92.5

sometimes afraid on the street			
Yes	9.5	10.8	12.2
No	90.5	89.2	87.8

1. The production process

The newly developed interview system of CentERdata uses the internet for interviewing. The respondents fill in their questionnaires using a website. This interview system makes it possible to follow the steps a respondent takes in real-time. As soon as they respond, their answer is known. The moment they encounter a problem, CentERdata is aware of it. This makes it possible to give online support during a questionnaire. CentERdata aims at a system where almost all of the problems are solved online. This system does not only save costs but is also a great step forward in supporting the respondent.

The system of weekly interviewing is more complex than one might think at first glance. Usually, each week more than one project is run. Sometimes, response is needed on questionnaires of previous weeks by those respondents who lagged behind. This requires an advanced selection system which assigns weekly the right questionnaires to the panel members. This weekly selection now is also made through the internet. The survey research division makes the selection by filling in a form on a website, thus deciding which respondent receives which questionnaires. The selection is saved in a selection database, where all the selections are stored. Some of the background information of the respondents is also stored in a database called the household database. The use of databases makes it possible to make more complex selections. Selections can for example been made on the base of response behaviour, formerly filled in questionnaires, geographical variables, income, number of children and so on. More complex selections can be made in cooperation with the system manager. All the selections (also the selection made by the survey research division on internet) are completed by using SQL-statements on the CentERdata databases.

The answers of the respondents are also stored in databases. The moment the respondent gives an answer to a questionnaire, the answer is stored in an database at CentERdata. This makes it possible to view the results of a questionnaire in real-time. These results can be published to the internet at the same time. It is possible to make a press release the weekend before the questionnaire is held. This press release should then contain a small explanation of the survey and an internet address of the place where the results can be found. CentERdata uses a dynamic way of publishing her information to the internet, so a graphical reflection of the collected data (collected so far) can be made. This way of publishing data is unique. The press or other interested parties can be made aware of the survey before the data collection has even started. They can follow in real time the results building up and use them when they are satisfied that a sufficient number of panel members has responded.

As mentioned before, CentERdata is concerned by the question of representativity. Not all respondents have an internet connection at their disposal. These respondents are provided with a so called settop box. This ‘box’ makes it possible to connect to the internet by using a telephone line, a television and a cordless keyboard. By pressing a button on this keyboard, the connection to the internet is made and the CentERdata questionnaire website appears. Some of the respondents may not even know they are connected to the internet. They are just filling in their questionnaire and after they completed their questionnaire, the settop box is shut down. There is no need to be familiar with the internet in any way to fill in a questionnaire.

2. The role of Blaise

CentERdata uses Blaise for her interviewing. The newly developed Blaise for Internet system makes it possible to make questionnaires through the internet. There are currently three options to fill in a questionnaire.

1. 1 The offline version. With this version a complete questionnaire can be made without a connection to the internet. A connection is made once, the questionnaire is loaded and the questionnaire can be made. After the questionnaire is completed, a connection to the internet is made to send the answers back.

2. The online version. By using this package, a respondent stays online with the internet. A more complex questionnaire can be made by using this version of Blaise for Internet.

3. The last version is a combination of the online and offline version. This version is cost efficient when large parts of the questionnaire are simple, but a few complex lookups or decisions have to be taken.

The HTML code the Blaise for Internet package delivers to a web browser is rather complex. CentERdata experienced a number of problems using the package on the newly bought settop boxes. For example, the Blaise for Internet package uses a version of javascript that the settop boxes couldn't understand. This problem has nothing to do with the quality of the Blaise package. The settop boxes are just not complex enough for the job. This, together with a problem concerning the identification of the respondent while filling in a questionnaire, made it necessary for CentERdata to develop a communication program in between the settop box and Blaise for Internet. The HTML code which is delivered to the communication program by Blaise for Internet is translated into HTML code which is understandable for the settop box and the information from the settop-box is made understandable for the Blaise for Internet package. Nothing was changed to the Blaise package, not by CentERdata and not by the CBS.

The layout of the questionnaire is very important to CentERdata. Our respondents are familiar with a certain layout. By the transition of the former interview system to the Blaise for Internet system, it was our goal to use a rather similar layout. The modelib editor makes it possible to design your own layout with a questionnaire. Due to the problems with the settop boxes and the communication program CentERdata had to build, all the layout elements are now handled in this program. The current layout is illustrated in the following figures

Figure 1. Layout of a SET-question

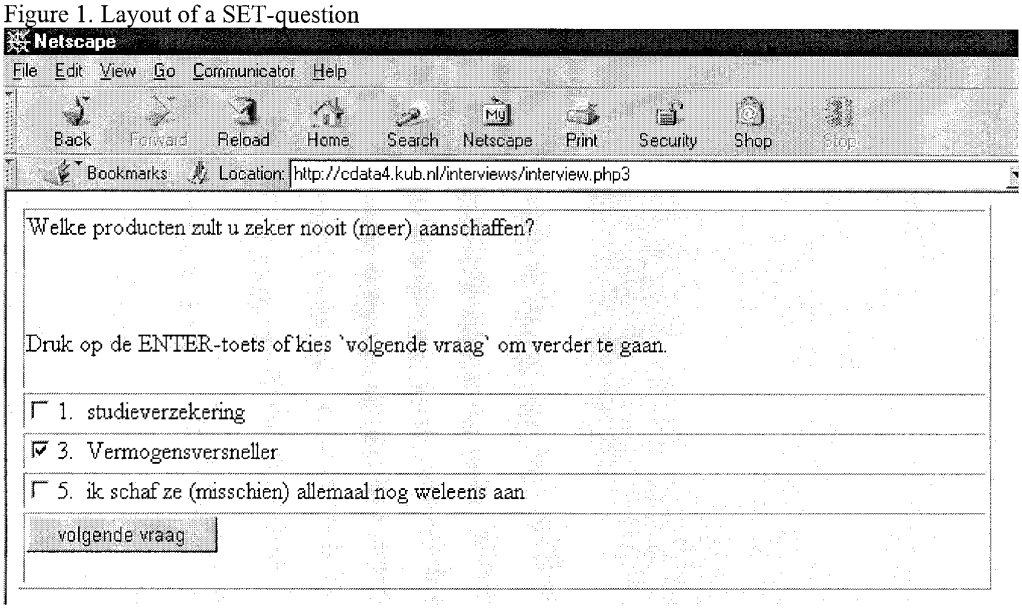
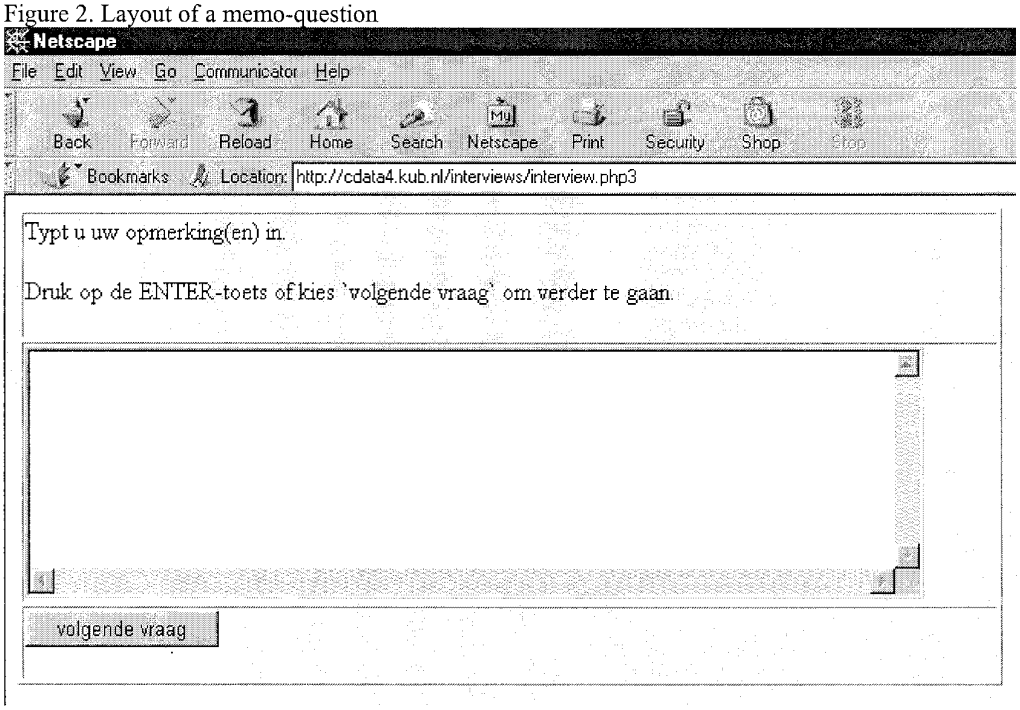


Figure 2. Layout of a memo-question



1. Two examples

5.1 The CentER Savings Survey

The CentER Savings Survey (in short: the CSS) is a panel survey that started in 1993. Each year, financial data are collected with 2000 households of the CentERpanel. The data contain information about work, pensions, accommodation, mortgages, income, assets, liabilities, health, perception of the personal financial situation, perception of risks, and much more. The data are unique because with these data it is possible to research both economic and psychological aspects of saving behavior. The CSS consists of five questionnaires:

Work and Pensions
Accommodation and Mortgages
Income and Health
Assets and Liabilities
Economic and Psychological Concepts

We will focus on the questionnaire on assets and liabilities, the most complex of these five. Assets and liabilities are measured in a very detailed way. Assets are investigated by distinguishing over 20 different asset components. Liabilities are divided into eight components (see table 2). In this questionnaire we assess the total value associated with each asset component, the (financial) institutions, the name of the product, the term, the interest rate, etc. The questions with respect to liabilities are similar.

Table 2. Overview of asset and liability components

assets	
checking accounts	put options bought
employer-sponsored savings plans	put options written
savings accounts linked to a Postbank account	call options bought
savings and deposit accounts	call options written
deposit books	real estate
savings certificates	cars
single-premium annuity insurance policies and annuities	motorbikes
savings or endowment insurance policies	boats
growth funds	caravans
mutual funds or mutual fund accounts	money lent out to family or friends
(mortgage) bonds	other assets
shares	
liabilities	
private loans	loans to family or friends
extended lines of credit	study loans
debts based on payment by installment etc.	credit card debts
debts with mail order firms etc.	debts not mentioned before

In order to improve the quality of the data collection we will pay attention to the following topics

1. the use of a help function to explain financial terms
2. providing overviews
3. using data gathered in earlier waves

5.1.1 Help function

The questionnaire on assets and liablitities contains many financial terms that require explanation to the respondents. Terms like ‘annuity insurance’, ‘single-premium insurance’, ‘endowment insurance policies’, ‘growth funds’ may not all be clear to the respondent. We use HTML anchor tags to provide the help texts. In the Bliase questionnaire we define *local variables* that we will use for *variable text fills*. The local variables contain anchor tags, so that the internet browser will show the financial term underlined, allowing the respondent to click on the financial term and get the help text. In Blaise we need the following code:

```
LOCALS
  TxtGrowthFund: STRING;
FIELDS
  Bank “With which bank or financial institution did you make the investment with your
^TxtGrowthFund?”:
    TBank
  Name “What is the name of your ^TxtGrowthFund?” : Tname;
  Value “How much was the value of the investment with your ^TxtGrowthFund on 31
December 1999?”:
    TEigth9s;
RULES
  TxtGrowthFund:= ‘<A HREF=css.php?state=help&item=‘‘growth fund’’>growth
fund</A>
  Bank
  Name
  Value
```

When filling out the questionnaire, the internet browser will display the text of the question *Bank* as: ‘*With which bank or financial institution did you make the investment with your growth fund?*’. If the respondent clicks the term ‘growth fund’ the internet browser will display the page: *css.php?state=help&item=‘growth fund’*. This page (HTML file) containing the help text is generated by PHP. The file ‘css.php’ may look as follows (we added some extra code for showing overviews which we will discuss later):

```
<?
$HelpArray = array();
```

```
$HelpArray['growth fund'] = "<B>Growth funds</B> are investment funds that do not pay
out interest or dividends, but invest their returns in the fund itself. In this way, no income tax
has to be paid on the returns.";
$HelpArray['annuity insurance'] = "By taking out <B>annuity insurance</B> the insured is



function ShowHelp($arg_item){
  global HelpArray;
  echo $HelpArray[$arg_item];
  echo "<BR><BR>\n";
  echo "<A HREF=javascript:history.back()>Return to the questionnaire</A>\n"
}

function ShowOverview(…){
  …
}

/* MAIN */
echo "<HTML>\n";
echo "<HEAD><TITLE>CentER Savings Survey</TITLE></HEAD>\n";
                    BGCOLOR=\"#D3F4E9\">\n";

switch ($state) {
  case "overview":
    ShowOverview(…);
  break;

  case "help":
    ShowHelp($item);
  break;
}
echo "</BODY>";
echo "</HTML>";
?>
```

5.1.2.Providing overviews

Filling out the questionnaire on assets and liabilities is not an easy task for a respondent. He may easily loose track of the assets that she already reported or did not yet report. It is therefore important to provide overviews. Providing overviews is desirable, but not a straightforward task in the internet version of Blaise. The usual way to construct overviews in

Blaise is to use table questions, but the table layout is not supported in the internet version. Therefore we use PHP in order to provide these overviews. In Blaise the code for a question *Check* that allows the respondent to get an overview, may look as follows:

```
FIELDS
 TableGrowthFunds: TTableGrowthFunds;

AUXFIELDS
 Check "Please click the button and check the growth funds that you reported!
        @/<FORM ACTION=css.php METHOD=post>
        @/<INPUT TYPE=hidden NAME=state VALUE= ""overview"">
        @/<INPUT TYPE=hidden NAME=asset_component VALUE= ""growth funds"">
        @/<INPUT TYPE=hidden NAME=column_header1 VALUE= ""bank"">
        @/<INPUT TYPE=hidden NAME=column_header2 VALUE= ""name"">
        @/<INPUT TYPE=hidden NAME=column_header3 VALUE= ""value"">
        @/<INPUT TYPE=hidden NAME=column1[] VALUE=
""TableGrowthFunds.Row[1].bank"">
         …
        @/<INPUT TYPE=hidden NAME=column1[] VALUE=
""TableGrowthFunds.Row[10].bank"">
        @/<INPUT TYPE=hidden NAME=column2[] VALUE=
""TableGrowthFunds.Row[1].name"">
         …
        @/<INPUT TYPE=hidden NAME=column2[] VALUE=
""TableGrowthFunds.Row[10].name"">
        @/<INPUT TYPE=hidden NAME=column3[] VALUE=
""TableGrowthFunds.Row[1].value"">

        @/<INPUT TYPE=hidden NAME=column3[] VALUE=
""TableGrowthFunds.Row[10].value"">
        @/<INPUT TYPE=submit  VALUE=Show overview>
        @/
        @/Is everything correct?": TYesNo;
```

In the internet browser the question *Check* will appear as:

Please click the button and check the growth funds that you reported!

Show overview

Is everything correct?

The action of showing the overview is laid down in the file ‘css.php’. We define the function ShowOverview as follows:

```
function ShowOverview($arg_asset_component,
   $arg_column_header1, $arg_column_header2, $arg_column_heade3,
   $arg_column1,$arg_column2 ,$arg_column3) {
  echo "These are the $arg_asset_component you reported:<BR>\n";
  echo "<TABLE>\n";
  echo "<TR><TD><B>$arg_column_header1</B></TD>”;
  echo "<TD><B>$arg_column_header2</B></TD>”;
  echo "<TD><B>$arg_column_header3</B></TD></TR>";
 $counter = 0;
 $endcounter = 0;
 while($counter < 10):
   if ($arg_column1[$counter] == ""):
   else:
     echo "<TR><TD>$arg_column1[$counter]</TD>";
     echo "<TD>$arg_column2[$counter]</TD>";
     echo "<TD>$arg_column3[$counter]</TD></TR>";
   endif;
   $counter += 1;
 endwhile;
 echo "</TABLE>\n";
 echo “<A HREF=javascript:history.back()>Return to the questionnaire</A>\n”
}
```

In the internet browser the overview of the growth funds will appear as:

These are the *growth funds* you reported:

bank	**name**	**value**
ABN-AMRO	All Dollar Bond Fund	12345
ING Bank	ING Bank Rente Groeifonds	67890

Return to the questionnaire

5.1.3. Using data that were gathered in an earlier wave.

The CSS is a panel survey where the five questonnaires are administered every year. It is our experience that the respondents do not find it easy to fill out the questionnaires. This is especially true for the questionnaire on assets and liabilities: there is a high burden in filling out all the detailed questions on all the different asset components. Things get even worse for those components that were reported earlier and have changed little (or not at all) since the last time. For those components it seems reasonable to provide the respondent with the data

that they reported earlier and ask till what extend the information changed. In order to include the previous data we read all the previous answers into the ‘external’ *PreviousData*. This is done as follows:

```
DATAMODEL Assets;
USES Assets;
PRIMARY Ident;
EXTERNALS PreviousData : Assets ('previousdata\assets'); {a local path that contains
previous data}
AUXFIELDS
  PreviousDataExists "Previous data available?" : TYesNo;
FIELDS
   Ident: TNine9s;
RULES
  Ident;
  PreviousDataExists.Keep;
  IF (Ident <>EMPTY) AND (PreviousDataExists = EMPTY) THEN
    IF PreviousData.Search(Ident) THEN
      PreviousDataExists:= yes;
      PreviousData.READ
   ELSE
      PreviousDataExists:= no;
    ENDIF
  ENDIF
  PreviousDataExists.SHOW
```

Having previous data at our disposal, we still have to decide on the amount of feedback we want to give the respondent, and on the reactions we allow the respondent to give. In the example below we provide the respondent all available information. Furthermore we allow the respondent either to change all answers, to only change the value of the growth fund, or to state that nothing has changed.

```
TYPE
  TStillOwn =
     Yes_nothing_changed  "Yes, and nothing changed",
     Yes_value_changed  "Yes,but the value changed",
     Yes_more_changed "Yes, but some properties changed",
     No "No, I don’t have it any more");
AUXFIELDS
  StillOwn   "Last time you reported a ^TxtGrowthFund called ^Name with the ^Bank, and you
reported
                  that its value was ^Value. Do you still own it?: TStillOwn;
```

```
FIELDS
  Bank "With which bank or financial institution did you make the investment with your
^TxtGrowthFund?":
    TBank
  Name "What is the name of your ^TxtGrowthFund?" : Tname;
  Value "How much was the value of the investment with your ^TxtGrowthFund on 31
December 1999?":
    TEigth9s;
RULES
  IF (PreviousDataExists=Yes)
  THEN
    Bank:= PreviousData.Bank
    Name:= PreviousData.Name
    Value:= PreviousData.Value
    StillOwn;
  ENDIF;
  IF (StillOwn=Yes_nothing_changed) OR (StillOwn=Yes_value_changed) THEN
     Bank.Show; Name.Show
  ELSEIF (StillOwn=Yes_more_Changed) OR (PreviousDataExists=no)  THEN
     Bank; Name
  ENDIF
  IF (StillOwn=Yes_nothing_changed) THEN
    Value.Show;
   ELSEIF (StillOwn IN [Yes_valueChanged, Yes_more_Changed]) OR
(PreviousDataExists=no)) THEN
    Value;
  ENDIF
```

5.2 The innovation survey

The innovation survey is an example of a set of processes which are followed on a monthly basis. Each process is the introduction of a new product to the market. Consumer behaviour is modeled in the following way. With regard to each of the products the consumer finds himself in one of the following states

0. never heard of the product
1. knows the name of the product and is able to classify the product
2. has some interest in the product (does not rule out the possibility that he may buy)
3. has knowledge of the product (and is able to compare with similar products)
4. has a positive intention to buy
5. has tried the product
6. has adopted the product (buys it on a regular basis); this does not apply to financial products
7. has rejected the product, will never buy it (again)

The flow of the meaasurement cycle is basically as follows:

- when measurements are made with respect to a new product, alle respondents start in the initial state 0 To determine whether the respondent knows a product the following question is asked

  ---

  Hear: "Below you see a list of product names. Of which of these product have you heard and do you know waht product it is? (don't click on it when you know the name but not the product type. You may click on more than one name": SET OF
  (a1 "Life Mortgage,
  a2 "Stock Mortgage",
  a3 "Combination Fund",
  a4 "Termijnkoopsompolis",
  a5 "Linear Mortgage",
  a6 "Education Insurance",
  a7 "Growth Power",
  a8 "Spavings Certificate",
  a9 "European Equity Fund",
  a10 "Clickfund",
  a11 "Saving Stocks",
  a12 "Capital Raiser",
  a13 "Pure Gold",
  a14 "Annuities Mortgage",
  a15 "Fortune Fund",
  a16 "NONE of the above");

  Some of the products are fake products; only three of them are of interest and cause follow up questions when clicked

- in each wave, the transitions from one wave to another are registered. For state 3, knowledge of the product, this question is

  ---

  Know "Do you know in what respects the Education Insurance is different from comparable products?": ARRAY [1..4] OF (no, yes)

- when a new state is entered, the respondents answers questions which are associated to this state. For state 5, trial, the following question is asked
  Buy "Is the Stock Mortgage only in your name or (also) in the name of someone else?": ARRAY[1..4] OF (a1 "only in my name", a2 "(also) in the name of someone else (spouse, family, friends)")

- for each of the products, the states in which the respondents finish the questionnaires are stored; in the next wave (usually one month later), the states are retrieved and serve as a starting point for the interview

```
USES WriteVar 'writevar';
EXTERNALS WriteFile: WriteVar('savenn');
FOR i:=1 TO 4 DO Mod1[i]:= WriteFile.Mod1[i] ENDDO;
FOR i:=1 TO 4 DO Mod2[i]:= WriteFile.Mod2[i] ENDDO;
FOR i:=1 TO 4 DO Mod3[i]:= WriteFile.Mod3[i] ENDDO;
FOR i:=1 TO 4 DO Mod4[i]:= WriteFile.Mod4[i] ENDDO;
FOR i:=1 TO 4 DO Mod4a[i]:= WriteFile.Mod4a[i] ENDDO;
FOR i:=1 TO 4 DO Mod5[i]:= WriteFile.Mod5[i] ENDDO;
FOR i:=1 TO 4 DO Mod6[i]:= WriteFile.Mod6[i] ENDDO;
FOR i:=1 TO 4 DO Mod7[i]:= WriteFile.Mod7[i] ENDDO;
FOR i:=1 TO 4 DO Mod8[i]:= WriteFile.Mod8[i] ENDDO;
```

- The states 6 and 7 serve as terminal states. When a respondent enters state 6 or 7 he will not be questioned about the product again. For financial products, state 5 replaces state 6 as terminal state.

The result of this procedure is a database in which invaluable information is stored about the course of different product innovations. It shows where problems arise in marketing, how communication in different media yields different results, which role the strengths and weaknesses of each product play and at what rates products are adopted by different target groups.

6. **Conclusion**

The combination of new the production system of CentERdata and the programming power of Blaise brings household data collection to the current internet standards, without loosing the long established scientific standards. Key ingredients are

1. a respresentative household panel
2. an interface between Blaise and HTML
3. easy database maintenance with the procedures that come with Blaise

It is to be expected that in the near future, when internet hardware is improved and new and better software is available, internet interviewing will become even more attractive, both for data collectors and respondents.